Masters Programmes

Assignment Cover Sheet

Submitted by:	**GroupA8 *<1955346, 1957541, 1958648, 1958943, 1980148,1992799>***
Date Sent:	**03/12/2019**
Module Title:	**Analytics in Practice**
Module Code:	**IB9BW0**
Date/Year of Module:	**2019**
Submission Deadline:	**04/12/2019**
Word Count:	**2199**
Number of Pages:	**9**

We are working for an analytics consultancy. A bank “Universal Credit” has approached the company to deliver a pitch to win a major contract with them to develop and deploy an improved direct marketing of a long-term deposit application with good interest rates. Several other consultancies have been approached as well, and the final decision on who will get the contract will largely depend on the outcome of a demonstration of the consultancy’s approach to this problem based on a dataset that the bank provided.

1. Introduction

Universal Credit Bank, a bank that provides financial products and services to individuals, is seeking ways to make their marketing strategies more effective. Currently, direct marketing via phone calls has been the sole strategy to reach out to customers to subscribe to their products or services, including the long-term deposit product. It is argued that the current direct marketing strategy is not efficient enough when it comes to targeting the right customers. The following statement summarises the business objective of this study:

> “Develop and deploy an improved direct marketing strategy for UCB for the long-term deposit product, with good interest rates, by following the Cross-Industry Standard Process for Data Mining (CRISP-DM) process.”

This report describes how the Cross-Industry Standard Process (CRISP-DM) can be used to answer UCB's case. First, a literature review of relevant topics on direct marketing is described. Then, the main findings on the application of data mining techniques for UCB’s case are outlined. Finally, a set of implications and recommendations for UCB is given in the conclusions section.

2. Literature Review

Data Mining for Direct Marketing

In a highly competitive banking sector, the most critical key to success is customer satisfaction. To achieve this, companies must offer campaigns that are personalised for customers. This is called direct marketing (Parlar and Acarvaci, 2017).

According to Ling and Li (1998), in comparison to mass marketing, direct marketing is primarily used by banking, retail and insurance sectors. Instead of doing a conventional analysis of customer patterns, it analyses each customer independently based on needs and characteristics. To examine this massive information, data mining techniques are used to create patterns to understand data and predict customer behaviour.

Selecting Informative Attributes

However, some problems might occur during the process of data mining. For example, extremely imbalanced class distribution when partitioning data; equally splitting testing and training sets; training set with significant variables could be too large for specific learning algorithms; predictive accuracy might not be a suitable evaluation measure and is weak for customer targeting as classification errors should be dealt with differently. To solve such issues, one of the approaches is selection of informative attributes (Ling and Li, 1998; Parlar and Acarvaci, 2017). As mentioned by Parlar and Acarvaci (2017), elimination of useless attributes and their sizes increases accuracy and efficiency. The most common attribute selection techniques are Information Gain (IG) and Chi-square test.

Sampling methods for Data Preparation

Marinakos and Daskalaki (2017) compare the impacts of different resampling methods applied to models. Those resampling methods, such as Distance-data resampling and a cluster-based under-sampling technique, were used to solve the imbalanced training sets and improve the performance of models. In our models, we also focus on bank direct marketing analysis and have a similar problem, that is one of the classes is relatively rare. Therefore, we try to use one of its resampling methods, Minority Oversampling Technique (SMOTE), to produce new synthetic observations randomly on the line segments that join any two neighbouring minor observations, and then compare with the method that we learned to see whether it can better solve the imbalanced training data and significantly improve model performance.

Modelling Methods

Elsalamony (2014) demonstrates the analysis and application of data mining techniques to achieve marketing strategies. The problem for this study was to identify the effectiveness of the campaign conducted by a Portuguese banking institution to attract customers for subscribing a term deposit. The primary objective of the paper is to assess the efficacy of the models used to predict the required results. The study was conducted using a multilayer perception neural network (MLPNN), tree augmented Naïve Bayes (TAN) known, Logistic regression (LR), and Ross Quinlan new decision tree model (C5.0). This closely relates to the objective of our goal.

The analysis uses statistical measures of accuracy, sensitivity and specificity and helps understand the data and draw connections to analyse the results of our model. The comparison of prediction and output, using the training and testing data based on the statistical measures for each model, gives us a good thought process of which model is suitable for the best results. As the paper uses the improvised versions of Naïve Bayes and Decision tree models, we have applied the experimental results to favour our predictions and analysis. The author conducts Naïve Bayes classification, as it not only classifies the algorithm but also utilises confidence measurements by ranking training and test examples, which in turn can be applied to our modelling process (Ling and Li, 1998).

Karim and Rahman (2013) investigated two data mining techniques: Naïve Bayes and C4.5 decision tree to summarise data for selecting a group of customers. The goal was to predict accurately whether a client would subscribe to a term deposit and find characteristics that mostly affected banks' profits by affecting a client's choice. Then banks can take steps to change the characteristics so that clients will fall into the desired node and become a term depositor.

The paper also extracted actionable knowledge from decision tree using a novel algorithm that obtained actions that were associated with attribute-value changes of clients from one status (not a depositor) to another (depositor). We have gained from this paper some inspirations on modelling. We found that to improve decision tree building and model evaluation, instead of the accuracy, area under the curve (AUC) of the ROC curve can be used to evaluate probability estimation. And attributes with the most subjects in the desired node are not necessarily the most profitable targets. Thus, we should always make a trade-off between traditional tree and profit optimal tree (SBP) based on our goal (Karim and Rahman, 2013).

Oslon and Chae (2012) compared variants of classical data mining techniques, including logistic regression, decision trees, and neural network against RFM (Recency, Frequency, Monetary) based predictive modelling. RFM relies on three customer behaviour variables to find valuable customers and develop future direct marketing campaigns (recency, frequency and monetary).

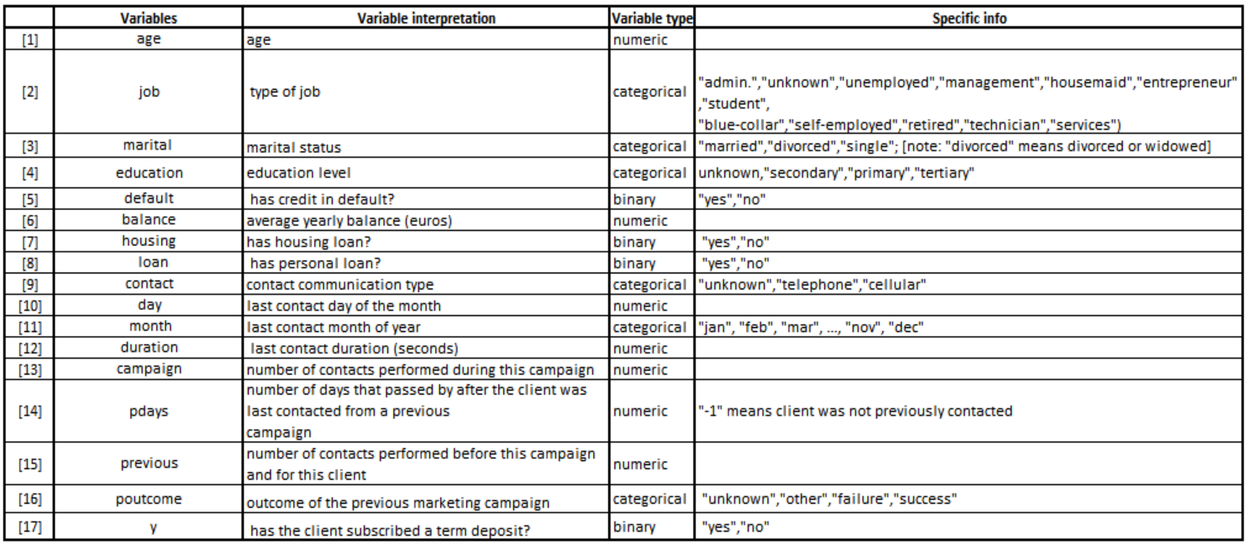
Although the RFM model can nicely help categorise customers based on their historical behaviours, the classical data mining techniques still outperform RFM models in terms of prediction accuracy and cumulative gain. The good performance of data mining methods indicates that RFM variables alone can be useful for building a reliable customer response model. However, since the classical data mining techniques have the advantage of adding other external variables in addition to R, F, M, all the three models performed better. Since the final call is with the marketing professionals, we should be aware of the trade-off between simple and sophisticated models and develop a well-balanced model using their domain expertise (Oslon and Chae, 2012).

3. Data Mining Process

3.1 Data Understanding

The dataset described in Table 1 includes 45,211 past UCB customer records based on 16 independent variables and one target variable "y", which demonstrates whether the customer has subscribed for the long-term deposit or not. There is no missing value in the dataset. However, the proportion of the target variable in the dataset is not balanced as 5289 customers have subscribed for the term deposit that states - ‘yes’ and 39922 records of customers who have not subscribed stating ‘no’. Below is the data dictionary:

Table 1 - Description about the dataset



	Variables	Variable interpretation	Variable type	Specific info
[1]	age	age	numeric	
[2]	job	type of job	categorical	"admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
[3]	marital	marital status	categorical	"married","divorced","single"; [note: "divorced" means divorced or widowed]
[4]	education	education level	categorical	unknown,"secondary","primary","tertiary"
[5]	default	has credit in default?	binary	"yes","no"
[6]	balance	average yearly balance (euros)	numeric	
[7]	housing	has housing loan?	binary	"yes","no"
[8]	loan	has personal loan?	binary	"yes","no"
[9]	contact	contact communication type	categorical	"unknown","telephone","cellular"
[10]	day	last contact day of the month	numeric	
[11]	month	last contact month of year	categorical	"jan", "feb", "mar", ..., "nov", "dec"
[12]	duration	last contact duration (seconds)	numeric	
[13]	campaign	number of contacts performed during this campaign	numeric	
[14]	pdays	number of days that passed by after the client was last contacted from a previous campaign	numeric	"-1" means client was not previously contacted
[15]	previous	number of contacts performed before this campaign and for this client	numeric	
[16]	poutcome	outcome of the previous marketing campaign	categorical	"unknown","other","failure","success"
[17]	y	has the client subscribed a term deposit?	binary	"yes","no"

3.2 Data Preparation

Before applying modelling techniques, data has to be prepared. First, all variables data types are reviewed and set to its supposed correct data type. Then, the information gain to the target variable for each attribute were calculated. Then, the dataset was partitioned into training and testing set by setting the ratio to 70% and 30% respectively. Sampling methods were used to balance the training set since the target variable has a class imbalance, which might result in a bias towards the majority class. SMOTE or Over-Under sampling methods (both) method were used against the training set. Finally, the balanced training set was sliced, having only the top 5 attributes with the highest information gain and the target variable.

3.3 Modelling

Since the objective is to classify whether a customer will subscribe to a term-deposit or not, the machine learning models selected have to support classification cases. Four widely known models were tested for this case, namely; Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and Random Forest. Decision tree and Naïve Bayes were chosen since both have performed well in Elsalamony (2014)_ while SVM and random forest are argued to be the industry's standard for classification.

To build a model that can fulfil the business objective, the data to be fed into the model has to be robust. By using the balanced and sliced training set, overfitting to the training set can be avoided. After fitting the model to the training set, the model was tested against the test set. The illustration of model fitting and model testing is comparable to a student training himself for the actual exam using mock exam questions. Once the models are built, the models can then be compared and evaluated.

3.4 Results and Evaluations

This subchapter describes the highlights from the modelling process by comparing the models' performance and contextualise the key findings to the business objective.

Informative Attributes

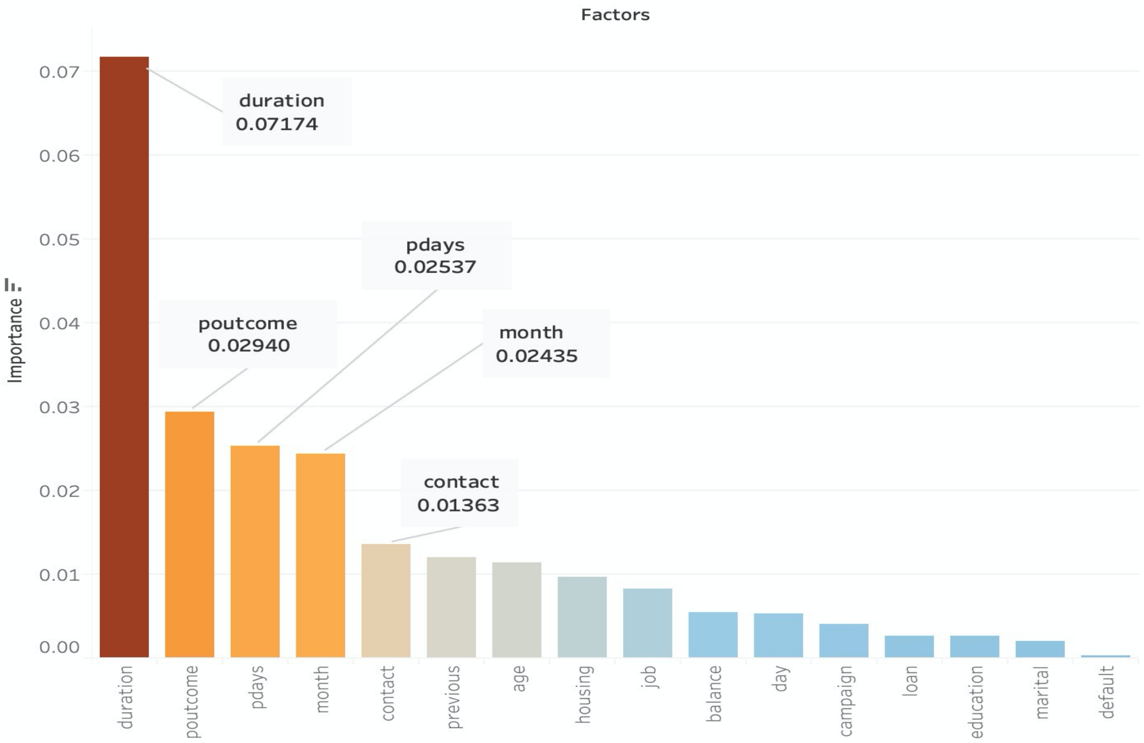


Figure 1 - Information gain chart

We utilised the information gain to reveal the importance of each attribute in terms of determining whether a customer will subscribe to the long-term deposit. Figure 1 illustrates the importance of each attribute out of 16 variables. From the graph, we can see the five most critical attributes are duration, poutcome, pdays, month and contract. It was found that “duration” could offer the most information for classifying potential customers.

Model Performance

Performance of each model was compared using five measures:

a) Accuracy ($\text{i.e.}\ \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$), measures the percentage of correct predictions based on total predictions of models. Random Forest has the highest accuracy, with 82.58% for successfully differentiating consumers who will subscribe term deposits from those who will not.

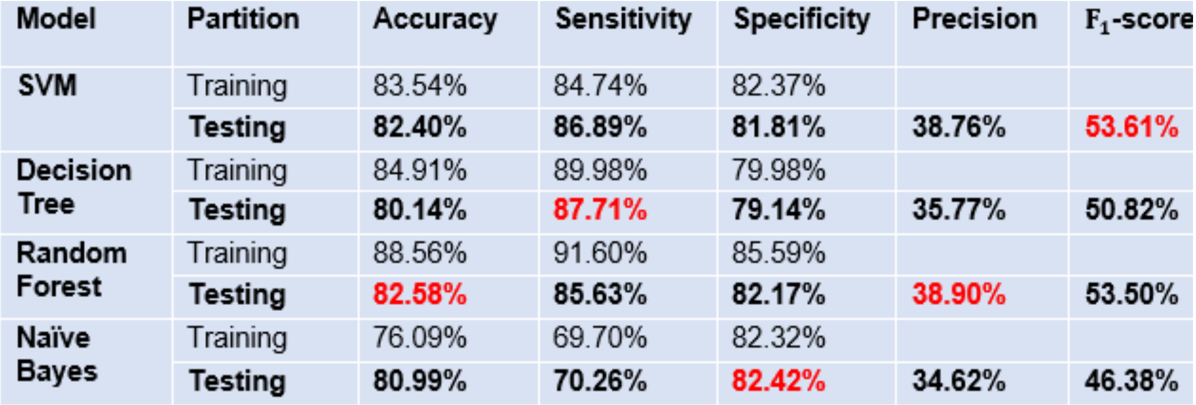
b) Sensitivity ($\text{i.e.}\ \frac{\text{TP}}{\text{TP+FN}}$), the measure of how the model can target all subscribed customers. Decision tree turns out to be the highest (82.58%).

c) Specificity ($\text{i.e.}\ \frac{\text{TN}}{\text{TN+FP}}$), the measure of how the model can classify the non-subscribers. Naïve Bayes has the highest score of 82.42%.

d) Precision ($\text{i.e.}\ \frac{\text{TP}}{\text{TP+FP}}$), the measure of true positive overall actual positives. The random forest has the highest precision of 38.90%.

e) F1-score ($\mathrm{i.e.}\ 2*\frac{\mathrm{Precision*Sensitivity}}{\mathrm{Precision+Sensitivity}}$), the measure of the balance between precision and sensitivity. When the bank plans to the best balance between sensitivity and precision rate, SVM with 53.61% F1-measure is the most optimal.

Table 2 - Model comparison



Model	Partition	Accuracy	Sensitivity	Specificity	Precision	F_1-score
SVM	Training	83.54%	84.74%	82.37%		
	Testing	**82.40%**	**86.89%**	**81.81%**	**38.76%**	**53.61%**
Decision Tree	Training	84.91%	89.98%	79.98%		
	Testing	**80.14%**	**87.71%**	**79.14%**	**35.77%**	**50.82%**
Random Forest	Training	88.56%	91.60%	85.59%		
	Testing	**82.58%**	**85.63%**	**82.17%**	**38.90%**	**53.50%**
Naïve Bayes	Training	76.09%	69.70%	82.32%		
	Testing	**80.99%**	**70.26%**	**82.42%**	**34.62%**	**46.38%**

ROC and AUC

A receiver operating characteristic curve (ROC) plots the Sensitivity and Specificity of a binary classifier as its discrimination threshold are differed. Figure 2 narrates the ROC chart for the four selected models; SVM, Random Forest, Naïve Bayes, and Decision Tree. The more the line aligns to the left of the graph, the better the model.

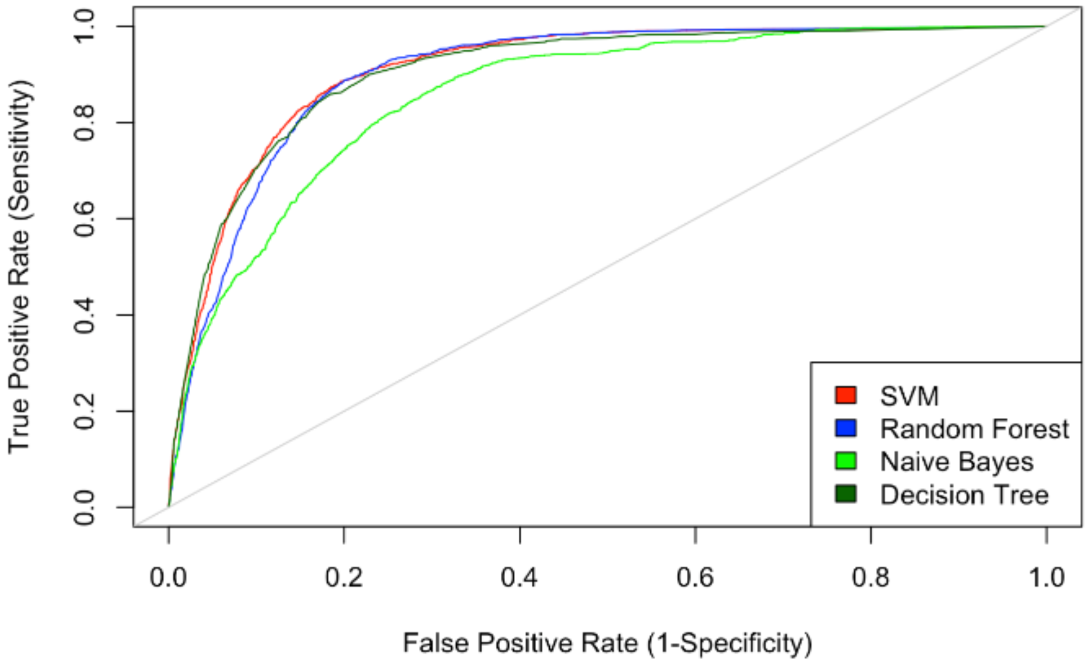


Figure 2 - ROC chart

Table 3 - AUC table

Model	SVM	Decision Tree	Random Forest	Naïve Bayes
AUC	0.9108	0.9054	0.903	0.8585

In order to find the model with best overall performance, AUC (the area under the curve of ROC), a statistical method used to describe the performance of each classifier, was utilised. The higher the AUC is, the better the performance of the classifier over different threshold points. Table 3 describes the AUC for each model. It can be concluded that SVM is the best overall model, while Naïve Bayes is the worst, relative to others.

<u>Gain Chart</u>

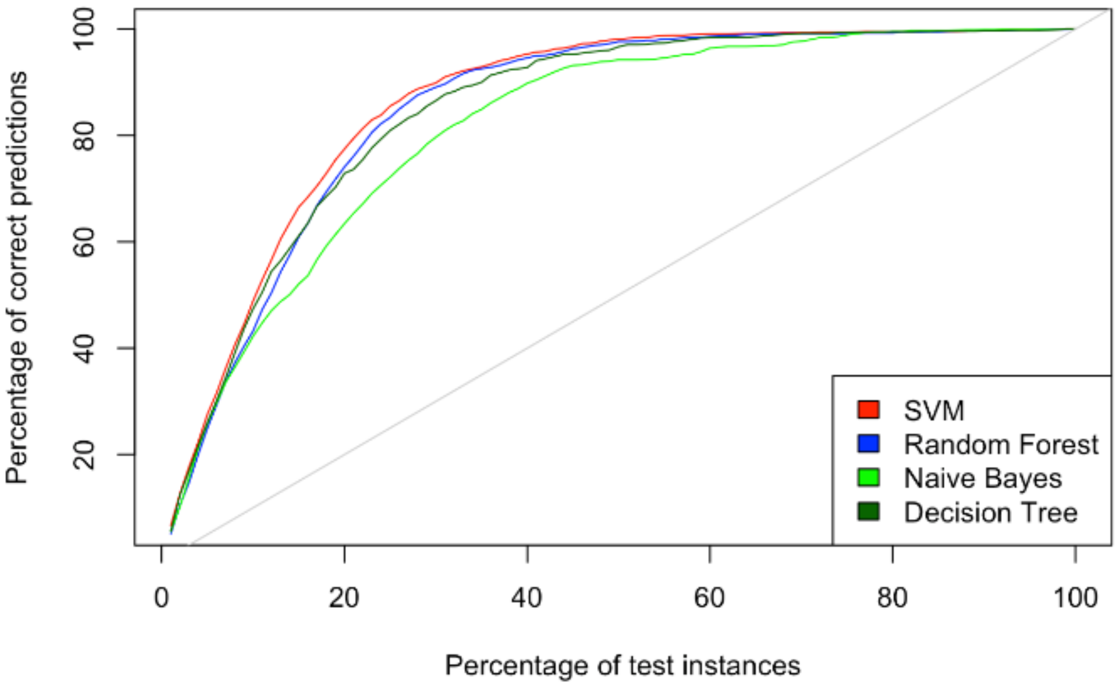


Figure 3 - Cumulative gain chart

The lines above evaluate the capability to state correct predictions and are ranked according to the probability of the outcome. For instance, by testing 20% of the data, the SVM model can achieve nearly 80% of correct predictions (i.e. customers response to campaign and subscription). However, compared to SVM, other three models require more samples to achieve accurate predictions.

4. Conclusions

This report is aimed to uncover ways for UCB to develop strategies to improve its current direct marketing strategy for its long-term deposit product by utilising data mining techniques.

UCB should start focusing its effort on the most influential variables that correspond highly to whether or not a customer will subscribe to its financial product. It was found that duration, a timeframe within which customers are kept interested about the product(long-term deposit) is a variable with the highest information gain. Therefore, to get a higher probability of convincing a customer to subscribe, employees' need to be trained with good communication skills. In addition, length and gap of contact also matters, so the bank should frequently update the customers about products and benefits.

After building and testing predictive models for UCB's case, Support Vector Machine (SVM) model stood out in terms of performance relative to other models introduced in this study. The SVM model is highly recommended to tackle UCB's case since it has the highest AUC value, highest gain chart position and F_1-score, along with relatively high accuracy and sensitivity rates. Therefore, by utilising SVM in direct marketing, UCB can expect to target the right customers for direct marketing campaign. However, clearly dependant on the marketing cost and expected return, other models might prove to have a better outcome. Therefore, it is recommended to roll out the model on a pilot project first before scaling up to the entire organisation.

References

Elsalamony, H. A. (2013) ‘Bank Direct Marketing Analysis of Data Mining Techniques’, *International Journal of Computer Applications*, 85 (7), pp. 12-22 [Online]. Available from: https://www.researchgate.net/publication/263054095_Bank_Direct_Marketing_Analysis_of_Data_Mining_Techniques (Accessed: 7 November 2019).

Karim, M. & Rahman, R. M. (2013) ‘Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing’, *Journal of Software Engineering and Applications*, 6 (4), pp. 196-206, SCIRP [Online]. Available from: https://www.scirp.org/journal/paperinformation.aspx?paperid=30463 (Accessed: 8 November 2019).

Ling, C. X. & Li, C. (1998) *Data Mining for Direct Marketing: Problems and Solutions*, The University of Western Ontario [Online]. Available from: https://www.aaai.org/Papers/KDD/1998/KDD98-011.pdf (Accessed: 10 November 2019).

Marinakos, G. & Daskalaki, S. (2017) ‘Imbalanced customer classification for bank direct marketing’, *Journal of Marketing Analytics*, 5 (1), pp. 14-30, SpingerLink [Online]. Available from: https://doi.org/10.1057/s41270-017-0013-7 (Accessed: 8 November 2019).

Olson, D. L. & Chae, B. K. (2012) ‘Direct marketing decision support through predictive customer response modeling’, *Decision Support Systems*, 54 (1), pp. 443-451, ScienceDirect [Online]. Available from: https://doi.org/10.1016/j.dss.2012.06.005 (Accessed: 11 November 2019).

Parlar, T. & Acaravci, S. K. (2017) ‘Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data’, *International Journal of Economics and Financial Issues*, *7*(2), pp.692-696 [Online]. Available from: https://dergipark.org.tr/en/download/article-file/365990 (Accessed: 11 November 2019)