WARWICK BUSINESS SCHOOL

Assignment Details:

Submitted by: 2227324, 2293968, 2222727, 2092171, 2247408, 2249760

Date Sent: 07/12/2022

Module Title: Analytics in Practice

Module Code: IB9BW0

Date/Year of Module: 2022

Submission Deadline: 08 Dec 2022, 12:00

Word Count: 1935

Number of Pages: 10

Question: *Data Mining for Email Marketing Campaign*

"I declare that this work is entirely my own in accordance with the University's Regulation 11 and the WBS guidelines on plagiarism and collusion. All external references and sources are clearly acknowledged and identified within the contents.

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done it may result in me being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work."

Table of Contents

Introduction

This report outlines all the analytical steps within the CRISP-DM methodology framework to perform the analysis for the Universal-Plus (referred to as 'Client'). In addition to Literature Review, this report is divided into six main sections for Data Mining steps, as listed below:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Model Evaluation
6. Deployment and Final Suggestions

Literature Review

In this paper, Ben-David (2008) discusses the relationship between ROC curves and Cohen's Kappa. The author points out that accuracy is the most common measure for assessing classifier accuracy, however it is problematic as well. This study proves pivotal for our current study as we have utilised ROC curves and Kappa values to evaluate the model performance of various classification models that we have implemented. In our study, we have eliminated Logistic Regression and Decision Trees based on the discussion of kappa values in this paper.

Ngai et al. (2009) examined the application of data mining techniques in customer relationship management in 87 selected papers. They also illustrated that the selected papers used various data mining techniques for direct marketing to achieve customer attraction, commonly adopted modelling techniques includes logistic regression, decision tree and clustering for decision making. This paper provided direction and a selection of various modelling techniques that we could adopt for our raw data, to achieve the goal of predicting which customers will visit the shop through direct email marketing in our project.

Huang et al. (2007) uses two credit datasets of UCI database to demonstrate the accuracy of SVM classifier. SVM based model can directly objectify the applications as rejected or accepted, reducing the risk of the creditor. It is at par with BPN and GP in terms of accuracy. This paper was selected to learn more about SVM machine learning technique. By comparing the paper's acceptance and rejection of applications model to our dataset, it gave us more insights and a direction to use SVM in an optimal way for predicting number of customers visiting through direct email marketing.

In the paper by Speiser et al. (2019), different variable selection techniques were assessed and compared in the setting of random forest classification based on different types of datasets. The prediction error rates, number of variables, computation times, area under the receiver operating curve, etc., were evaluated. The study concluded that the method implemented in the R packages of VSURF, varSelRF, and Boruta were the best variable selection methods for most datasets, which provided an insightful direction for our approach of using Random Forest to interpret and analyze customer behaviors.

Ullah et al. (2019) used different machine learning algorithms to predict churn customers, including logistic regression, decision trees, SVM and random forests. These models are evaluated using performance metrics like accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area. It is observed that Random Forest algorithm produced the best result for the above customer churn prediction. This paper is inspirational to our project as it is highly related to our work, in terms of context, possible models, and evaluation metrics. The result of the authors' work is also a strong hint that random forest model might be the best candidate for building the customers classification model required by our client.

Saito and Rehmsmeier (2015) suggest that for imbalanced datasets, ROC plots might not be a good performance illustrator because it is not sensitive to false positive rate. While precision will drop drastically with high false positive rate and can reflect the performance of the model on imbalanced datasets more informatively. This paper is inspiring for our project as the dataset for our assignment is also imbalanced, with 10,181 entries being positive and 53,819 entries being negative. At the same time, we are also more concerned about false positive rate, as our main objective is to minimize the cost of targeting uninterested customers.

Data Mining

Section 1 – Business Understanding:

1.1 Background:

We are a part of an Analytical consulting company, and our client has requested us to design and deploy an e-mail marketing system to target customers and increase sales.

1.2 Problem Statement:

Presently, the Client is following rules of thumb or randomly targeting their customers. Given the present random approach of targeting customers, some of their previous target campaigns were not successful.

1.3 Objective:

To design a methodology to predict the target customers who will visit the shop because of the direct e-mail campaign utilizing the CRISP-DM methodology. The constraint is that targeting uninterested customers costs the company money. In other words, we need to minimize the 'False Positives' to achieve the objective.

Section 2 – Data Understanding:

The input data provided by the client has 64,000 customer records contained in a .CSV file; each customer is associated with 20 uniquely identifiable attributes. Below are the observations from the initial data exploration:

1. The input data does not contain any duplicates at the customer level. Therefore, no de-duping was done. Also, data can be considered fairly clean and robust for further steps.
2. There are missing values in two attributes – 'purchase_segment' and 'spend' which are dealt in the data preparation step.
3. The given data set is imbalanced with respect to the target variable i.e., 'visit'. Out of 64,000 unique records, only 10,181 (~ 16%) instances have positive class. This suggests that, during the data preparation step, we need to utilize the sampling techniques on the training data to get the better precision rate.

Section 3 – Data Preparation:

Below data preparation steps are done after importing the input data provided by the client in the r environment:

1. An appropriate data type correction process is done to avoid any potential error in the model building step. For instance, attributes like 'new_customer' and target variable 'visit' are converted to factor data type from the integer data type.
2. There are 26 missing values in the 'purchase_segment' attribute. However, these missing values are re-calculated based on the 'purchase' attribute (purchase_segment_v1) during the data preparation step, as there are no missing values in the 'purchase' attribute.
3. Similarly, the 'spend' column has 49 missing values. At this stage, we cannot omit these records because for all such records, the customer has visited. Therefore, these records are essential for the model-building process. Instead, missing value treatment is done for these records by replacing the missing values with the mean of the 'spend' where the customer has visited (because 'spend' is simply 0 for customers without a visit).
4. Visual inspection of attributes like 'customer_id' and 'account' suggests that these

columns are not appropriate for further steps in CRISP-DM methodology as ‘customer_id’ is unique for each record, and ‘account’ is always 1.

5. As shown below in Table 1, the ‘age’ attribute has 0 information gain for the target variable. However, converting age into segments (age_segment) like ’19-30’,’30-40’,’40-50’, and ‘50+’ gives positive information gain and makes more business sense. Therefore, instead of the ‘age’ attribute, we have retained the ‘age_segment’ attribute for the model building.

   On the other hand, information gain for the ‘spend’ attribute is the highest. However, out of 64,000 data observations, 53,819 observations have 0 spend value and did not visit (visit = 0). Given, our data is imbalanced, in other words, most customers did not visit the store, therefore, despite having higher information gain, the 'spend' feature is not a good predictor for modeling.

Table 1 – Percentage Information Gain for Target Variable Visit

Attribute	Information Gain (%)
Spend	83.361
recency	9.596
email_segment	3.538
purchase	1.102
purchase_segment_v1	1.082
channel	0.479
delivery	0.340
womens	0.217
new_customer	0.121
zip_area	0.047
Phone	0.039
Mens	0.035
payment_card	0.033
employed	0.004
marriage	0.003
dependent	0.001
age_segment	0.001
Age	0.000

Section 4 – Modelling:

The train/test split of 70/30 is utilized for model building and evaluation process respectively. Also, as discussed in Section 2 - Data Understanding, we have utilized the oversampling method to handle the class imbalance, maintaining a 50% class ratio in the training data from the rare class. Further, this oversampled data is the input data for all the model-building steps.

Below mentioned are the supervised classification models that were implemented to meet the objective of the problem. Please refer to Section 5 for the discussion on the reasons which is the best models to deploy.

1. Logistic Regression (LR)
2. Linear Discriminant Analysis (LDA)
3. Support Vector Machine (SVM)
4. Decision Tree (DT)
5. Random Forest (RF)

Section 5 – Model Evaluation:

Each of the models mentioned in Section 4 is critically evaluated according to the objective that targeting uninterested customers costs the company money. In other words, we need to minimize the 'False Positives' or maximize precision to achieve the objective.

We have utilized an elimination process to ultimately select the best model out of the five models mentioned above. Below mentioned are the steps followed to arrive at the best model:

Step 1 – As shown in Table 2 – Summary of Accuracy & False Positive Instances, we were able to eliminate the LDA model for the reasons mentioned below:

1.1 Maximum number of False Positive Instances (5,251), and

1.2 Least Accuracy of all five models (68.49%)

Given the stringent range of Accuracy and False Positive Instances, the rest four models need additional measure to conclude the best model.

Table 2 – Summary of Accuracy & False Positive Instances

Model	Accuracy %	False Positives Instances
LR	78.34	2,902
LDA	**68.49**	**5,251**
SVM	78.66	3,300
DT	78.78	2,788
RF	88.68	779

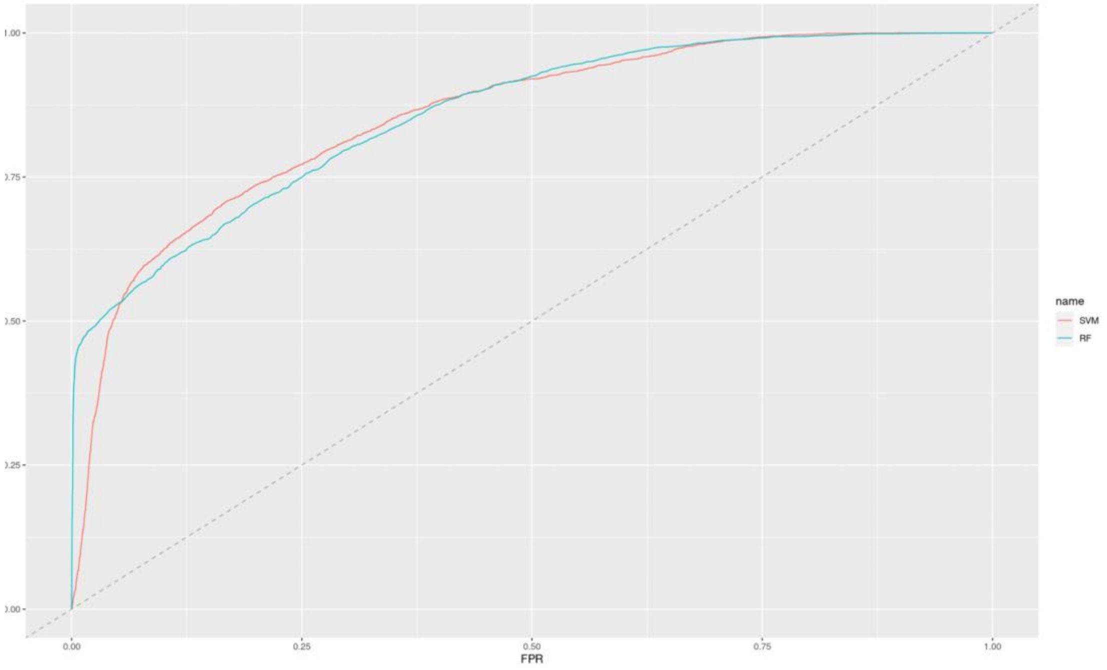
Step 2 – As shown in Table 3 – Summary of Accuracy, f1 Scores, and Kappa values, we were able to eliminate LR and DT models based on a comparison between f1 scores and kappa values. f1 score of LR (46.36) and DT (46.44) suggests that they are weaker compared to SVM (52.42) and RF (59.21); also, kappa values for LR (0.3354) and DT (0.3385) indicate poor fit. Hence, it is safe to conclude that based on the given input data, any model based on these techniques may not perform well during the deployment stage. Note at this stage, the f1 scores and kappa values for SVM and RF are comparable (given – accuracy alone is not the sole determining criteria), we can't conclude which of SVM or RF is better.

Table 3 – Summary of Accuracy, f1 Scores, and Kappa Values

Model	Accuracy %	f1 score	kappa values
LR	**78.34**	**46.36**	**0.3354**
SVM	80.01	52.42	0.4013
DT	**78.78**	**46.44**	**0.3385**
RF	88.68	59.21	0.5259

Step 3 – In this step, an attempt has been made to compare SVM and RF based on ROC curves and AUC values. Below Figure 1 – ROC curve for SVM (AUC = 0.8543) & RF (AUC = 0.856) shows that none of the models can be conclusively eliminated based on ROC curves and AUC values.

Figure 1 - ROC curve for SVM & RF



Step 4 – Ultimately, based on the precision values (RF – 67.42 % > SVM – 40.62 %), we conclude that RF is a better model than SVM for future e-mail campaigns.

Section 6 - Deployment and final suggestions

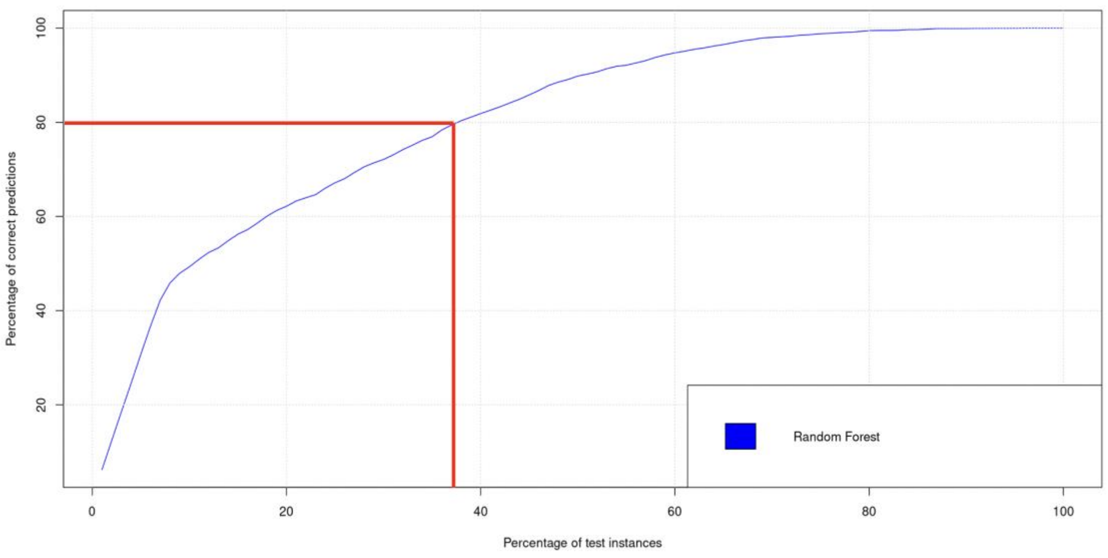
Based on the input data provided by the client and the objective of the study, our analysis suggests that the Random Forest algorithm (Precision ~ 67%) with the below-mentioned technical parameters (Table 5 – Technical Parameters for Random Forest Model) could be deployed for the future e-mail campaigns.

Table 5 – Technical Parameters for Random Forest Model

Parameter	**Value**
Train/Test split	0.7
Oversampling from rare class	0.5
Number of trees to grow	600
Variables randomly selected as candidates at each split	6

As shown below in Figure 2 - Cumulative Gain chart for Random Forest, using ~40% of the data gives around 80% of the total positives.

Figure 2 – Cumulative Gain chart for Random Forest



Reference

Ben-David, A. (2008). About the relationship between ROC curves and Cohen's kappa. Engineering Applications of Artificial Intelligence, 21(6), pp.874–882. doi:10.1016/j.engappai.2007.09.009.

Huang, C.-L., Chen, M.-C. and Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), pp.847–856. doi:10.1016/j.eswa.2006.07.007.

Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, [online] 36(2), pp.2592–2602. doi:10.1016/j.eswa.2008.02.021.

Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), p.e0118432. doi:10.1371/journal.pone.0118432.

Speiser, J.L., Miller, M.E., Tooze, J. and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, pp.93–101. doi:10.1016/j.eswa.2019.05.028.

Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access, 7, pp.60134–60149. doi:10.1109/access.2019.2914999.