WARWICK BUSINESS SCHOOL

Masters Programmes
Assignment Cover Sheet

Submitted by: 2153008, 2288991, 2237004, 2229140, 2293446, 2216101

Group Number: 5

Date: 8 December 2022

Module Title: Analytics in Practice

Module Code: IB9BW0

Date/Year of Module: 2022

Submission Deadline: 8 December 2022

Word Count: 2080

Number of Pages: 18

Question:
Technical Report on Data Mining Project for Customer Visitation Prediction in E-mail 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."

Contents

1. Introduction

The necessity of targeting customers for efficient marketing campaigns has increased significantly. Market competition has affected lower customer responses and drives more expenses in a company (Chakraborty et al. 2014). This following report will explain the development of our predictive analytics models of customer response requested by Universal Plus. The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is used to explain the detailed steps which could propose reliable solutions to the problem.

2. Academic Review

The goal of the academic review phase was to understand the application impact of predictive modelling in direct marketing. According to the paper of Kumar (2020), data mining would increase profitability of companies and affect marketing strategies. The efficiency of the marketing system can be reached by analysing data more efficiently in each six different phases of the CRISP-DM process (Ozyirmidokuz et al., 2015). Data preparation phase is crucial part, without any effort to handle the imbalanced dataset, the model will always lead to biassed results as it will tend to misclassify the test data as the dominant class from the paper of (Bach et al., 2022). Crone et al. (2006) discussed that data preparation contributed to a significant to a significant 50-70% difference in predictive accuracy and outperformed. Based on the paper of Chae & Olson (2012), the important variables to process are the last purchase, how often the customer purchases, and how much the customer has bought. In the modelling phase, SVM, Decision Tree, and Random Forest had a different result (Kayes et al., 2019).

3. Problem Statement and Approach

3.1. Business Problem Statement

The business problem is an unsuccessful email marketing campaign that targets uninterested customers and creates wasted costs for the company. In competitive consumer markets, targeting interested customers play a significant role for the success of marketing campaigns. We are asked to implement a direct marketing system which can identify the target group and predict which customer will visit the shop as a result of the direct email campaign. Our solutions would allow Universal

Plus to take next steps to target the right customers and reduce the cost of marketing campaigns.

3.2. Data Understanding

Our team used the data collected by our client, Universal Plus, for all of their customers during a period of two weeks following the email campaign. The dataset contained of 64,000 instances of customer's behaviour information in the past time. The performance of existing marketing framework was really not effective, proven by facts that only 13.7% of total population of customers have received the marketing campaign and positively impacted to visit the shop. On the contrary, 53.4% of the total population did not visit the shop even though they were given e-mail marketing (Appendix 8.2.1). This finding displays that there was a huge wasted marketing cost invested to uninterested customers.

3.3. Data Mining Approach

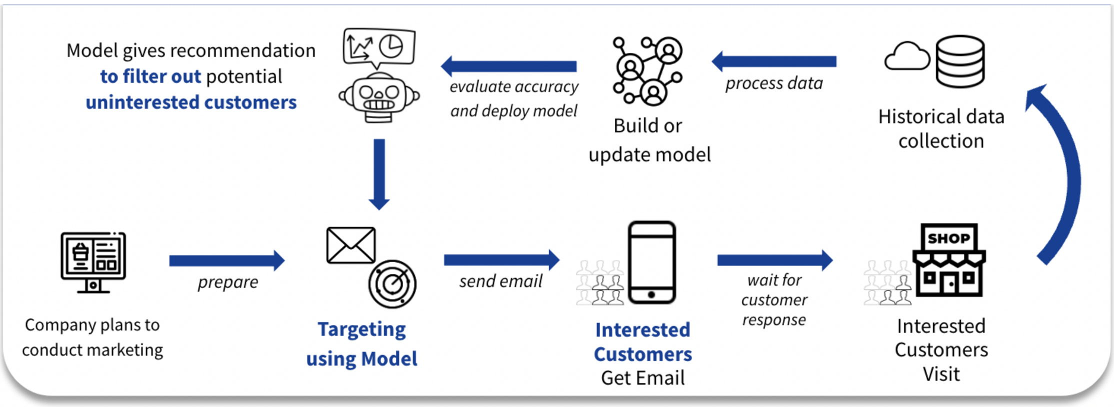


Figure 1. The illustration of proposed framework using prediction model

This report proposed to utilise a model prediction to help the business identify any potential uninterested customers, hence reducing the cost of marketing. The model depends on the past customer's behaviour data to make predictions. Several additional processes will be also added to enhance the existing marketing framework, for instance the process to collect historical data and new targeting mechanism.

The dataset had 20 attributes (Appendix 8.1), mainly grouped into 3 categories, such as purchasing history data, email segment, and customer demographics, which were used to predict customer's response to the marketing campaign. Only responses from customers, who have received campaigns, were taken as a valid dataset in building this prediction model. Data preparation process is done in building the model, as not every information is relevant and fits effectively in the model. For instance, high level of class imbalance, missing values and categorical variables, which existed in the dataset (Appendix 8.2.2 and Appendix 8.2.3), were treated using several pre-processing techniques.

Finally, we built the model with various types of classifiers to ensure the dataset to be trained through various algorithms. Therefore, we used five different approaches to proceed with and evaluate the performance of each approach to find out which classifiers give us the best prediction score in order to maximise cost minimization, which will be explained in the following sections.

4. Data Preparation

4.1. Importing and Cleaning data

Initially, a comma-separated values (CSV) file has been imported to examine data. Summary of the data showed that data types of some columns are not correct, and some columns do not give significant information about the target variable. For instance, the *Customer_ID* variable does not give information about the target variable. The *account* variable has only one level of information across the dataset. The *purchase_segment* column provides the same information with the *purchase* variable. That is reason of taking *Customer_ID*, *account* and *purchase_segment* variables out of the training and test dataset.

In order to include categorical variables in analysis we need to define the data type of categorical variables as factors. Hence, the data type of target variable *visit* and all other categorical variables have been converted from numeric into factor.

We built a model to predict visitation based on the impact of sending email to the customer, hence customers who are not given any email have been removed from the training and test dataset.

4.2. Handling with missing values

Missing values show that there is no input for some rows. When we examined the data set, we observed that the spend column had missing inputs. There are different methods to handle missing data points. In the spend column we replace missing values with the median of the column. Compared to the mean, median is not sensitive to extreme values. In order not to skew values in the spend column we used median value.

4.3. Data Normalisation

The pre-processing of the data is a crucial step which affects performance of the classification models. Singh, D. & Singh, B. (2020) discuss in their paper that data normalisation is essential in terms of transformation of the features. They outline in the paper that it helps reduce dominance of larger numeric values on smaller numeric values. As a result all features contribute to the learning process equally. In order to minimise the scaling difference between features we have applied min-max normalisation method on numeric futures.

4.4. Feature Selection

Feature selection is the process to pick informative variables that contribute to the learning process of the model. In this process insignificant features which do not contribute to the learning process are excluded from the model. We produced information gain chart (Appendix 8.2.4) to understand the effectiveness of the variables in the model. The higher information gain score the higher power the features have to contribute to the model.

To check multicollinearity between the features, we made a correlation matrix (Appendix 8.2.5). As a result of the correlation we see that *visit* and *spend* variables are very highly correlated. That is why we exclude the *spend* variable from the training and test dataset.

4.5. Data Partition

Our dataset is imbalanced. In order to give more information to the training set we decided the partition rate 80% for the training dataset and 20% for the test dataset. As our target is imbalanced we applied both under and over sampling methods to create balance between minor and major classes.

5. Modelling

5.1. Logistic Regression (using GLM)

Logistic regression (LR) is one of the classification models we used in our project. LR uses a logistic function to compute probabilities for the classes (Kayes et al., 2019). In this model, we also used ten-fold cross validation to get best result. Constantin (2015) states in his research that LR is a strong tool to analyse marketing strategies. As our target variable is binary, we could use LR to predict visitation.

5.2. Support Vector Machine

Support Vector Machine (SVM) is another widely used classification model. SVM uses information from the variables and creates appropriate classes using hyperplanes. The SVM classification model has multiple applications such as text recognition, disease diagnosis etc. It is also used to increase efficiency of marketing strategies (Kayes et al., 2019). That is why we applied SVM to predict the visitation.

5.3. Decision Tree

Decision tree (DT) method is used to measure the effectiveness of marketing strategies by many researchers. DT calculates information gain and entropy to test each node in the tree (Kayes et al., 2019). Liu (2022) used DT to analyse effectiveness of content marketing strategy. He states that among other classification models, DT gave more satisfactory results for content marketing analysis. We ran a model on a decision tree in our work to predict visitation.

5.4. Random Forest

Random forest (RF) is an ensemble machine learning algorithm which uses multiple algorithms (Kayes et al., 2019). Combining multiple algorithms makes RF a strong predictor of our analysis. Gao and Ding (2022) also constructed a digital marketing recommendation model using this algorithm and concluded that it has various advantages in terms of smaller generalisation error, more efficient high-dimensional data processing, fast training process, and better accuracy score.

5.5. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a decision tree based algorithm and it can be used to improve the performance of the models, especially for handling sparse data (Chen and Guestrin, 2016). Sparse data means that many of the values are recorded as zero, which is also quite similar to the structure of some features in the dataset. Mushava and Murray (2022) also proposed an classification approach using XGBoost of class-imbalanced data.

6. Evaluation

The goal of the model evaluation phase was to choose the best-performing predictive model that can predict which customers will visit the shop as a result of a direct email campaign. According to the paper of Muramira & Nkurunziza (2021) discussed the confusion matrix to evaluate the model performance.

We summarized our various approach's result (Appendix 8.3.1 - Appendix 8.3.7) into Table 1 and we decided that the best model was XGBoost which had dominant values in accuracy, area under the receiver operating characteristic (ROC) curve (AUC), fallout, and precision. According to the paper of Muramira & Nkurunziza (2021), precision and AUC are more considered to assess the effectiveness and the consistency of the model. Based on the five models we used, XGBoost's precision value was 83.10% which had the highest value among other models and had the highest number of correct visit predictions and the lowest number of false visit predictions. Additionally, based on the paper of Allaire (2006), AUC value can be categorised based on model performance. Besides that, XGBoost's AUC value was 87.89% which could be categorised as a very good model.

Model Classifier	Accuracy	Recall	Precision	AUC	Estimated cost reduction (in £)
Logistic Regression (GLM)	74.64%	64.75%	42.24%	76.48%	£ 67,959
SVM	77.34%	55.73%	45.65%	76.00%	£ 58,803
Decision Tree	83.17%	59.61%	58.81%	78.81%	£ 67,802
Random Forest	85.48%	64.80%	64.51%	87.45%	£ 81,692
XGBoost	88.80%	56.93%	83.10%	87.89%	£ 92,306

Table 1. Summary table of important metrics

Besides technical evaluations, we used cost reduction evaluation between an implementation without a model and a solution with a predictive model. The cost estimation was calculated with assumption of cost of email marketing per customer equals to £ 3. We calculated that marketing cost will only incur from the portion of True Positive (TP) and False Positive (FP) cases as we expected they will be only the customers who will be given email marketing. As the result, five models had a cost reduction ranging from £58,803 to £92,306 with the highest cost reduction being the XGBoost Model (Appendix 8.3.8).

7. Conclusion

In this study, five predictive models were proposed to identify the target group and predict which customers will visit the shop in order to target the right customers and reduce the cost of the marketing campaign. The best overall predictive model was obtained using the XGBoost model which outperformed in all metrics such as accuracy, fallout, precision, and ROC-AUC. This model also showed a significant difference value among other models in fallout and precision which resulted in a higher number of correct visit predictions. Additionally, the business evaluation of cost reduction between an implementation without a model and a solution with a predictive model was £92,306 for XGBoost. Based on those evaluations, the results of our study prove that the solution with our model is beneficial for our potential client Universal Plus.

8. Appendices

8.1. Dataset Description

Attribute Name	Attribute Description
Customer_ID	Customer identification number
recency	months since last purchase before the marketing campaign.
purchase	actual purchase in the past year before the marketing campaign.
purchase_segment	Categorisation for the purchase amount in the past year before the marketing campaign
mens	Whether the customer purchased men's merchandise in the past year before the marketing campaign.
womens	whether the customer purchased women's merchandise in the past year before the marketing campaign.
zip_area	categorisation of zip code as Urban, Suburban, or Rural.
new_customer	whether the customer is new in the past year or s/he is an existing customer.
channel	categorisation of the channels the customer purchased from in the past year.
email_segment	email campaign the customer received.
age	age of the customer in years.
dependent	whether the customer is dependent or not.
account	whether the customer has an account or not.
employed	whether the customer has a permanent job.
phone	whether the customer registered his/her phone or not.
delivery	categorisation for the delivery address.
marriage	marital status.

payment_card	whether the customer registered a credit card for payment in the past year
spend	total amount spent in the following two weeks period
visit	the flag to show whether the customer visited the shop in the following two weeks period or not

8.2. Exploratory Data Analysis

8.2.1. Total Customers based on Visitation and Email Treatment

	Visit		
Email Sent	Yes	No	Grand Total
Yes	8,764	34,017	**42,781**
No	1,417	19,802	**21,219**
Grand Total	**10,181**	**53,819**	**64,000**

8.2.2. The Proportion of Total Customers based on Visitation

Visit	Total Customers	% of Total Customers
Yes	8,764	20.49%
No	34,017	79.51%
Grand Total	**42,781**	**100.00%**

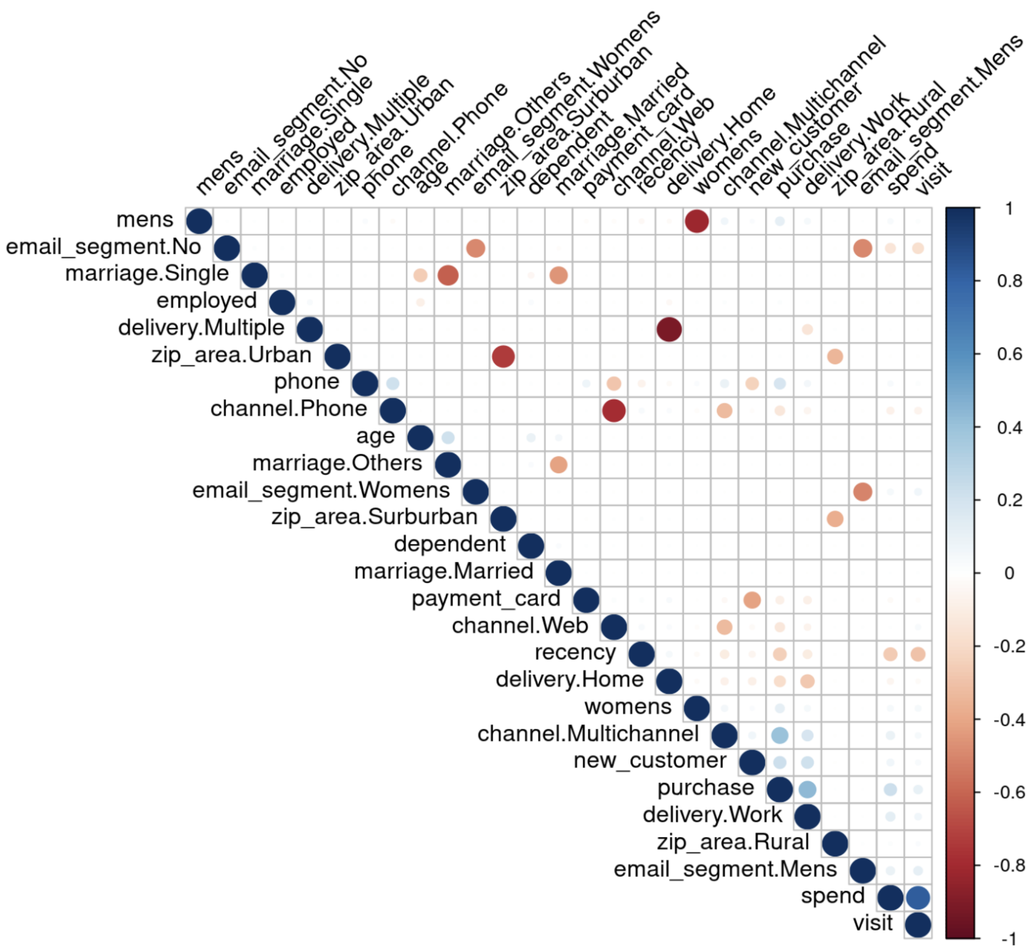
8.2.3. Missing values on Variable "Spend"

Spend Segments	Total Customers	% of Total Customers
Null	49	0.11%
£ 0 - £ 100	39,532	92.41%
£ 100 - £ 200	3,132	7.32%
£ 200 - £ 300	36	0.08%
£ 300 - £ 400	20	0.05%
> £ 400	12	0.03%
Grand Total	**42,781**	**100.00%**

8.2.4. Feature Importances in Table

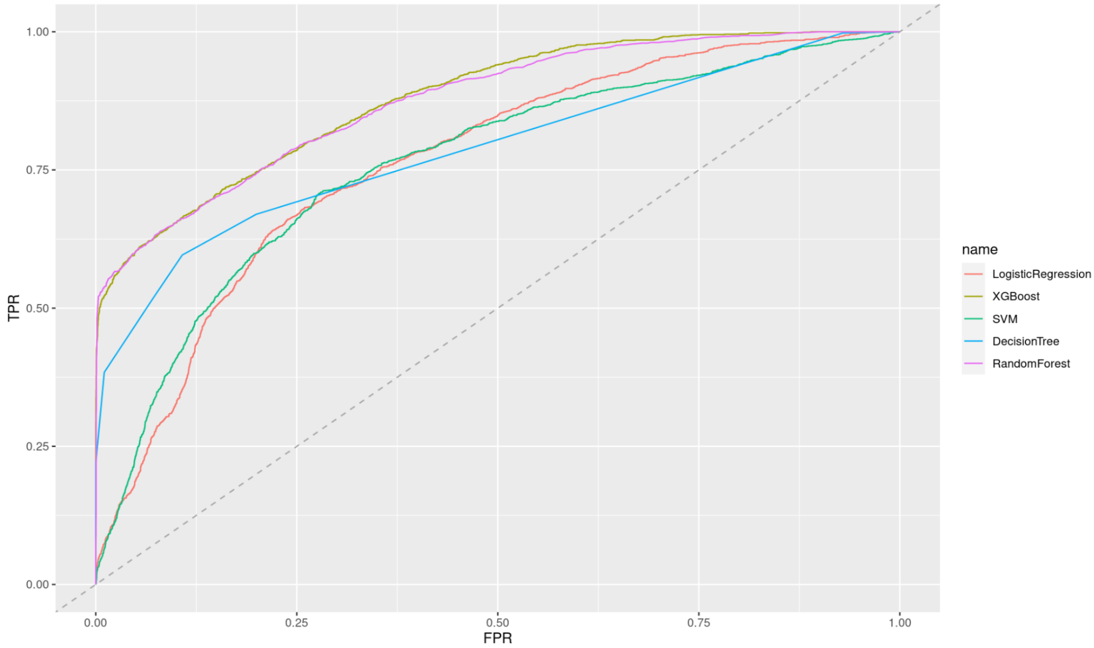
	attr_importance <dbl>
recency	4.926252e-02
email_segment	1.816504e-02
purchase	5.655904e-03
channel	2.458106e-03
delivery	1.746895e-03
womens	1.113348e-03
new_customer	6.190637e-04
zip_area	2.413959e-04
marriage	1.528872e-05
mens	0.000000e+00
age	0.000000e+00
dependent	0.000000e+00
employed	0.000000e+00
phone	0.000000e+00
payment_card	0.000000e+00

8.2.5. Matrix correlation chart (insignificant correlations are leaved blank)

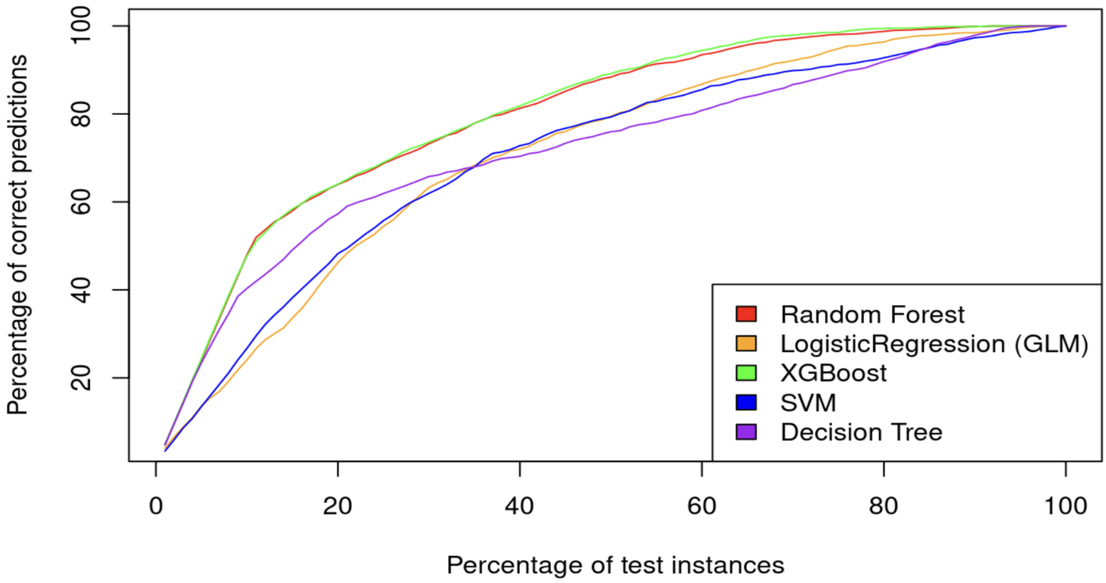


8.3. Evaluation Results

8.3.1. ROC Chart



8.3.2. Gain Chart



8.3.3. Logistic Regression (using GLM) Confusion Matrix Result

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5251  618
         1 1552 1135

               Accuracy : 0.7464
                 95% CI : (0.737, 0.7556)
    No Information Rate : 0.7951
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3501

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.4224
                 Recall : 0.6475
                     F1 : 0.5113
             Prevalence : 0.2049
         Detection Rate : 0.1327
   Detection Prevalence : 0.3140
      Balanced Accuracy : 0.7097

       'Positive' Class : 1
```

8.3.4. Support Vector Machine (SVM) Confusion Matrix Result

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5321  399
         1 1482 1354

               Accuracy : 0.7802
                 95% CI : (0.7712, 0.7889)
    No Information Rate : 0.7951
    P-Value [Acc > NIR] : 0.9997

                  Kappa : 0.4511

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.4774
                 Recall : 0.7724
                     F1 : 0.5901
             Prevalence : 0.2049
         Detection Rate : 0.1583
   Detection Prevalence : 0.3315
      Balanced Accuracy : 0.7773

       'Positive' Class : 1
```

8.3.5. Decision Tree Confusion Matrix Result

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6071  708
         1  732 1045
                                          
               Accuracy : 0.8317          
                 95% CI : (0.8236, 0.8396)
    No Information Rate : 0.7951          
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.4861          
                                          
 Mcnemar's Test P-Value : 0.5444          
                                          
              Precision : 0.5881          
                 Recall : 0.5961          
                     F1 : 0.5921          
             Prevalence : 0.2049          
         Detection Rate : 0.1221          
   Detection Prevalence : 0.2077          
      Balanced Accuracy : 0.7443          
                                          
       'Positive' Class : 1               
```

8.3.6. Random Forest Confusion Matrix Result

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6178  617
         1  625 1136
                                          
               Accuracy : 0.8548          
                 95% CI : (0.8472, 0.8622)
    No Information Rate : 0.7951          
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.5552          
                                          
 Mcnemar's Test P-Value : 0.8426          
                                          
              Precision : 0.6451          
                 Recall : 0.6480          
                     F1 : 0.6466          
             Prevalence : 0.2049          
         Detection Rate : 0.1328          
   Detection Prevalence : 0.2058          
      Balanced Accuracy : 0.7781          
                                          
       'Positive' Class : 1               
```

8.3.7. XGBoost Confusion Matrix Result

```
Confusion Matrix and Statistics


prediction_XGB    0    1
             0 6600  755
             1  203  998

               Accuracy : 0.888
                 95% CI : (0.8812, 0.8946)
    No Information Rate : 0.7951
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6109

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.8310
                 Recall : 0.5693
                     F1 : 0.6757
             Prevalence : 0.2049
         Detection Rate : 0.1166
   Detection Prevalence : 0.1404
      Balanced Accuracy : 0.7697

       'Positive' Class : 1
```

8.3.8. Estimated Cost Reduction Calculation

Model Classifier	% Total Customer								Cost Reduction Estimate		Total Cost		Cost Reduction	Expected Spend	Profit
	Random Email				Using Model										
	TN	FN	TP	FP	TN	FN	TP	FP	TP	FP	Without Model	With Model			
GLM					61,4%	7,2%	13,3%	18,1%	£834	£67.125		£60.297	£67.959		£90.449
XGBoost					73,7%	7,6%	12,9%	5,9%	£1.597	£90.709		£35.950	£92.306		£114.797
SVM	30,9%	2,2%	13,7%	53,1%	59,8%	4,1%	16,4%	19,8%	-£5.225	£64.028	£128.256	£69.453	£58.803	£150.746	£81.293
Decision Tree					61,9%	6,6%	13,9%	17,6%	-£310	£68.112		£60.454	£67.802		£90.292
Random Forest					69,4%	6,4%	14,1%	10,1%	-£804	£82.496		£46.564	£81.692		£104.182

9. References

Alitouche, T. A., Bekkar, M., & Djemaa H. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications,* 3, 10.

Allaire (2006). *Introduction à l'analyse ROC Receiver Operating Characteristic*. Centre de recherche Institut Philippe-Pinel de Montréal, école d'été.

Bach, M. P., Rogic, S., & Kascelan, L. (2022). Customer Response Model in Direct Marketing: Solving the Problem of Unbalanced Dataset with a Balanced Support Vector Machine. *J. Theor. Appl. Electron. Commer. Res.* 17(3), 1003-1018.

Batista, G., Prati, R. C., & Monard, M. C., (2004). A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data. *Special issue on learning from imbalanced datasets,* 6, 1.

Chae, B. & Olson, D. L. (2012). Direct Marketing Decision Support through Predictive Customer Response Modeling. *Decision Support Systems*, 54*,* 443-451.

Chakraborty, G., Mandapaka, A. K., & Kushwah, A. S. (2014). *Role of Customer Response Models in Customer Solicitation Center's Direct Marketing Campaign*. Oklahoma State University: Stillwater, OK, USA; pp.1–12.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

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*,* 781-800.

Constantin, C. (2015). Using the Logistic Regression Model in Supporting Decisions of Establishing Marketing Strategies . *Economic Sciences,* 8(57)*,* 2.

Forouzandeh, S., Sheikahmadi, A., & Soltanpanah, H. (2014). Content Marketing through Data Mining on Facebook Social Network. *Wobology,* 11, 1.

Gao, W. and Ding, Z., 2022. Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm. *Security and Communication Networks*, 2022.

He, H., Jadhav, S., & Jenkins, K. (2018). Information Gain Directed Genetic Algorithm Wrapper Feature Selection for Credit Rating. *Applied Soft Computing,* 69*,* 541-553.

Kayes, A. S. M., Sarker, I. H., & Watters, P. (2019). Effectiveness Analysis of Machine Learning Classification Models for Predicting Personalised Context-Aware Smartphone Usage. *Journal of Big Data volume 6*, Article number: 57.

Kumar, S. (2020). Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm. *Journal of Artificial Intelligence and Capsule Networks,* 02, 185-193.

Liu, Y., & Yang, S. (2022). Application of Decision Tree-Based Classification Algorithm on Content Marketing. *Journal of Mathematics,* 2022.

Muramira, A. & Nkurunziza, J. (2021). A Data-Driven Model to Predict a Household's Capacity to Graduate. *African Center of Excellence in Data Science,* 1-23.

Mushava, J. & Murray, M. (2022). A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, Volume 202.

Ozyirmidokuz, E. K., Uyar, K., & Ozyirmidokuz, M. H. (2015). A Data Mining Based Approach to a Firm's Marketing Channel. *Procedia Economics and Finance 27,* 77-84.

Singh, D. & Singh, B. (2020). Investigating the Impact of Data Normalisation on Classification Performance. *Applied Soft Computing,* 97*,* Part B.