Masters Programmes

Assignment Cover Sheet

Submitted by: 1914784, 2283394, 2283834, 2286979, 2288495, 2291942

Date Sent: 7 December 2022

Module Title: Analytics in Practice

Module Code: IB9BW0

Date/Year of Module: Term one 2022 - 2023

Submission Deadline: 8 December 2022

Word Count: 2176

Number of Pages: 14

"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

1. Introduction **1**

2. Literature Review **1**

3. Business Understanding **3**

4. Data Understanding **3**

5. Data Cleaning and Preparation **4**

6. Modelling **5**

6.1 Customers' visit prediction 5

1. Logistic Regression (LogReg) 5

2. Decision Tree (DT) 5

3. Random Forest (RF) 5

4. Support Vector Machine (SVM) 5

6.2 Customers' spend prediction 5

7. Evaluation **6**

8. Deployment **10**

9. Conclusion **10**

References **11**

Appendix A **12**

1. Introduction

E-mail marketing is one of the most effective ways to communicate with customers. This method is said to be cost-effective, easy to implement and profitable; however, it also requires the right approach to target the right customer to send the e-mail to obtain the best result from the e-mail marketing campaign (Grimes, Hough and Signorella, 2007).

This report will demonstrate an implemented e-mail personalised marketing system using CRISP-DM methodology to specify the targeted customer for the company. More explicitly, the report will outline how data mining can determine the right target customers based on their visits as a result of e-mail marketing. Additionally, it will analyse whether predicting customer spending can be used to determine the target group and calculate the trade-off between the cost and benefit of sending e-mails.

Figure 1. CRISP-DM methodology

2. Literature Review

A paper released in 2015 addressed the issue of unbalanced data when developing biometric models. Data imbalance determines the models to perform poorly, as the models can become biassed towards the majority class. This is especially detrimental when the negative class is the majority class and therefore the models fail to determine the patterns for the minority class. In order to improve the model's performance, they analysed hybrid sampling, which combines both undersampling (randomly removing attributes from the majority class) and oversampling (randomly duplicating attributes from the minority class).

The results showed that using hybrid sampling highly improved the model's performance and determined a better prediction for the positive class in terms of true positive rates. (Gazzah, Hechkel and Essoukri Ben Amara, 2015)

In 2007, McCarty and Hastak conducted research to observe how data mining improves targeting the right customers in direct marketing, using different models. They stated that the LogReg differs from other models as it can provide a probability for each individual customer in terms of response to the marketing campaign. Moreover, LogReg tends to perform really well when the relation between the predictor variables and the target variable is not continuous (McCarty and Hastak, 2007).

In 2015, Nachev and Teodosiev analysed how SVM performs on direct marketing campaigns. The benefit of using SVM is that it is suited to work with datasets with a large number of attributes, but they may not perform the best with datasets that have an extremely high number of observations because of the complexity of training the model. The Gaussian RBF kernel method has been proven to perform better than its alternatives in high input dimension dataset, including the linear and polynomial method (Nachev and Teodosiev, 2015).

In 2012, a paper on comparative analysis of decision trees and random forest modelling of breast cancer dataset with variation of instances yields different results for each model. Comparison of precision, recall and accuracy conclude that REPTree using informative gain (IG) as splitting criteria is more suitable for smaller data sets. Whereas, Random Forest performs effective discriminative classification as a result of ensemble strategies and random sampling, overcomes overfitting and is less sensitive to outliers. Hence, random forest produces precise and accurate results with larger data sets even when the number of instances are kept consistent (Ali et al., 2012)

There are several criteria that are used to evaluate the performance of modelling techniques. A 2021 study about churn prediction in digital games used confusion metrics which are outputs of a model showing True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The combination of them can determine the performance of a data mining classification model as it can be transformed into accuracy, recall, precision, F-score and, importantly, the ROC-AUC curve which are all the ratios of the confusion matrix. In order to have a good model, accuracy and recall should be as high as possible (Jain, Khunteta and Srivastava, 2021).

In 2008, Cui et al. analysed the performance of multiple models used for direct marketing, which was relatively difficult as there are trade-offs between sensitivity and

specificity. The decision of the best model was based on the AUROC analysis, summarising the TP rate (sensitivity) and FP rate (1 – specificity) by plotting a curve in a two-dimensional plane. AUROC significantly helped with the measurement of such trade-offs and the selection of the optimal threshold value that minimises misclassifications. In general, a model with a higher AUROC statistic is said to outperform the alternatives or dominate the others (Cui et al., 2008).

3. Business Understanding

“Universal Plus" aims to use direct e-mail marketing to reach out to their customers and increase sales. However, by randomly selecting customers for the campaign, the company is losing money from sending emails to uninterested customers. Therefore, the success of the campaign is determined by the precision of choosing the right customers to send e-mail; this involves determining their target customers which are likely to visit as a result of the marketing campaign, as well as avoiding e-mailing the customers that are unlikely to visit.

4. Data Understanding

The data used for the analysis provided by Universal Plus includes the details of customers that were part of a previous campaign. The dataset has 64,000 entries and 20 attributes, either numeric or characters (see more in appendix A). However, most attributes can be considered categorical variables, and there are no duplicated entries. The attributes *purchase_segment* and spend have NA values. Also, some attributes have a clear high correlation between them: *{purchase_segment: purchase}* and *{spend: visit}*. This is because *purchase_segment* only places into segments the amounts in purchase, while spend is the outcome of the visit. Therefore, it cannot be used to predict a visit.

The target variable is in the form of a binary output which indicates whether the customer visited after the marketing campaign or not. The Exploratory Data Analysis (EDA) determined some important relationships between attributes and the target variable. For instance, the more recently a customer purchased before the marketing campaign, the more likely they are to visit the store after the campaign (Figure 2).

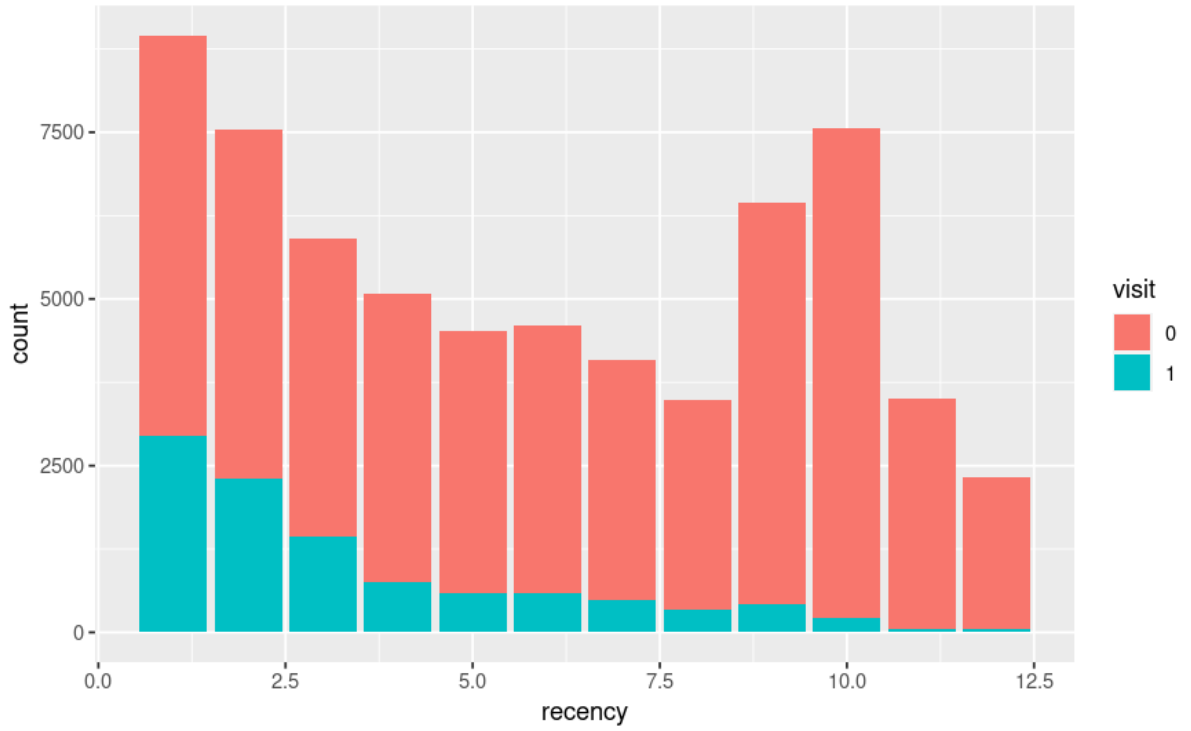

Figure 2. The relationship between attribute “recency” and the output variable “visit”

5. Data Cleaning and Preparation

The first step is to transform all the categorical variables and the target variable into factors. Then, a copy of the dataset is created, *df_emailed*, which stores only the records of the customers that received an email to analyse the customer’s visit behaviour as a result of receiving a marketing email. Further cleaning is done for modelling: remove the insignificant attributes (*Customer_ID*, *account*), the attributes from the pairs with high correlation (purchase_segment, spend) and adjust the levels for the *email_segment*. After cleaning, *df_emailed* is split into training and test, using a 0.7 split ratio. Because in the initial dataset the data is very unbalanced (approx. ratio 80:20), the training dataset is balanced by using both over and undersampling in order to prevent the model from being biassed towards the dominant class.

Another copy of the initial dataset is made, *df_spend*, in which the *spend* attribute is the target variable. The subset uses only customers who received an email and that have *spend* values non-negative. This is because only 75 customers out of 8764 don’t spend after they visit, which is insignificant. *Customer_ID, account, purchase_segment, visit* are removed for modelling and then *df_spend* is split into training and test with a 0.7 split ratio.

6. Modelling

6.1 Customers' visit prediction

1. Logistic Regression (LogReg)

Two LogReg models were developed, one with the unbalanced training set and another with the balanced training set. This was done to confirm the choice of hybrid sampling based on the data used for modelling, which is shown in the evaluation part. Once we confirmed that using balanced data improves the model's performance, the rest of the models were built using only the balanced dataset *training_both.*

2. Decision Tree (DT)

IG was used to determine the attributes with a high correlation with the target variable based on reducing the weighted average variance (entropy) for each individual attribute in relation to the target variable. Then, the top 10 attributes with the best IG were added to a new subset dataset, *subset_training*, to predict precise results. 2 DT models were trained with the balanced training set by using all the attributes, as well as using the top 10 subset dataset. Then, both DT models were tested on the same test set.

3. Random Forest (RF)

RF model works by generating multiple decision trees. The model was built with the balanced dataset for better precision and to overcome the overfitting possibility.

4. Support Vector Machine (SVM)

The SVM model is used as it is robust in mapping variables to predict accurate decisions in classification problems (Brerefton and Lloyd, 2010). The RBF kernel method is employed for the SVM model training.

6.2 Customers' spend prediction

To predict how much people would spend, two regression models were used: linear regression and a nonlinear regression model implemented with RF. The two models were used to decide which linear or nonlinear function would better describe the data patterns.

7. Evaluation

Evaluation of Classification for Predicting Customer Visit

For LogReg, the results of the model using a balanced training dataset are better compared to those of the model using the original training dataset, as shown in *Figure 3*.

For DT, the ROC curve was used to compare the performance of the two models (*Figure 4*). The graph can be interpreted as the model trained with the top 10 IG attributes (red line) performing slightly better than the one with all attributes (blue line).

Therefore, the LogReg with a balanced training dataset and DT model with the top 10 IG variables were chosen for the final evaluation between all four different models.

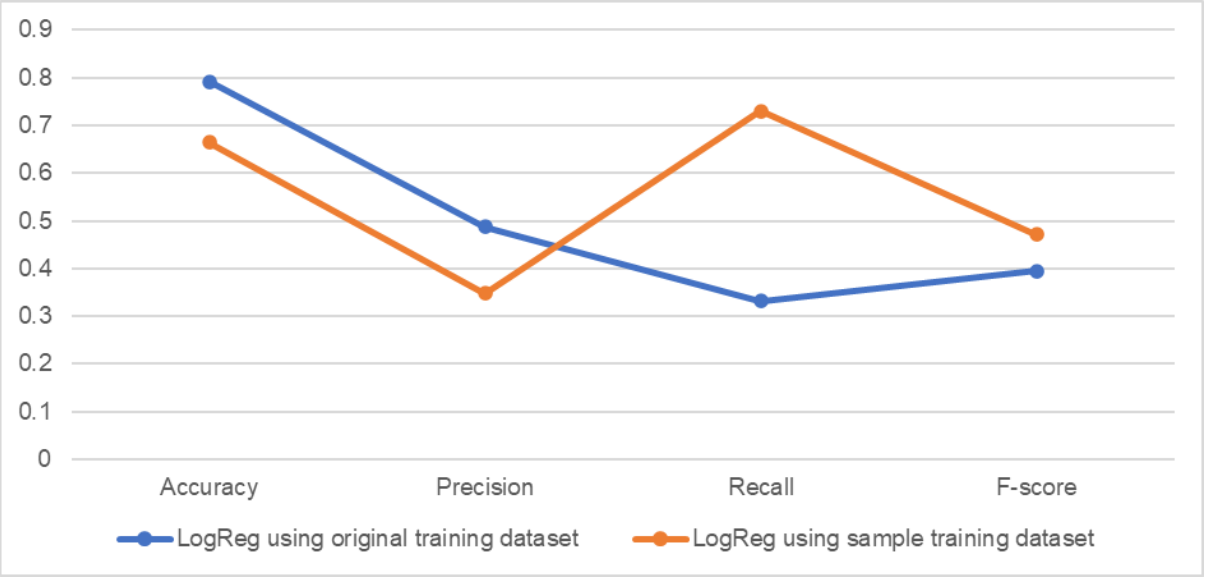

Figure 3. Graphic comparison of performance metrics for LogReg

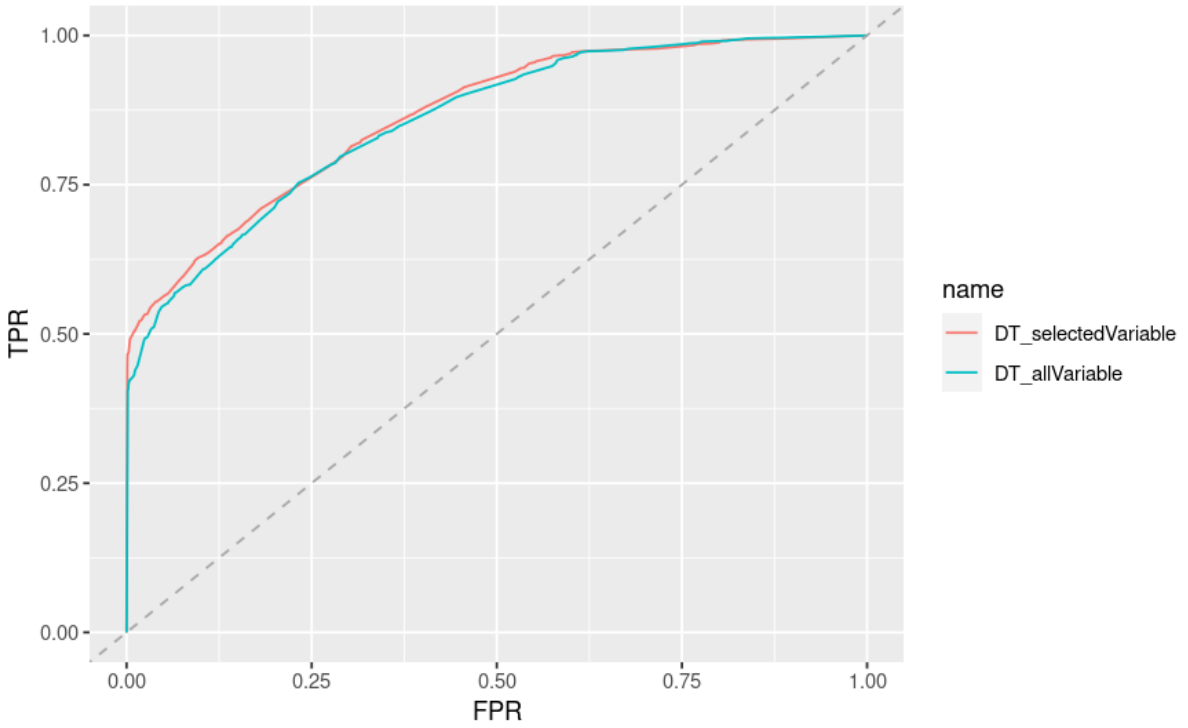

Figure 4. ROC curve compared DT models

The comparison of the performance metrics for LogReg, DT, RF and SVM is illustrated in *Figure 5*. The ROC curve was used to compare all four models (*Figure 6*), and the AUC value is shown in *Table 1*. *Figure 5* shows that RF performed best in accuracy, precision and F-score. Although it gave the worst recall, the difference percentage between RF recall and the model with the highest recall, SVM, was not significantly different, which was only 8.33% difference. Moreover, even though the AUC value of DT was the highest, followed by RF, the precision, which is the number of correct position predictions out of all positive predictions of RF, was higher. More specifically, a lower precision implies a higher cost for the company, while a higher recall represents a higher number of target customers. Therefore, by analysing the trade-off between the costs of sending emails and the benefits of having a high number of target customers, it was concluded that the best model for the business is RF.

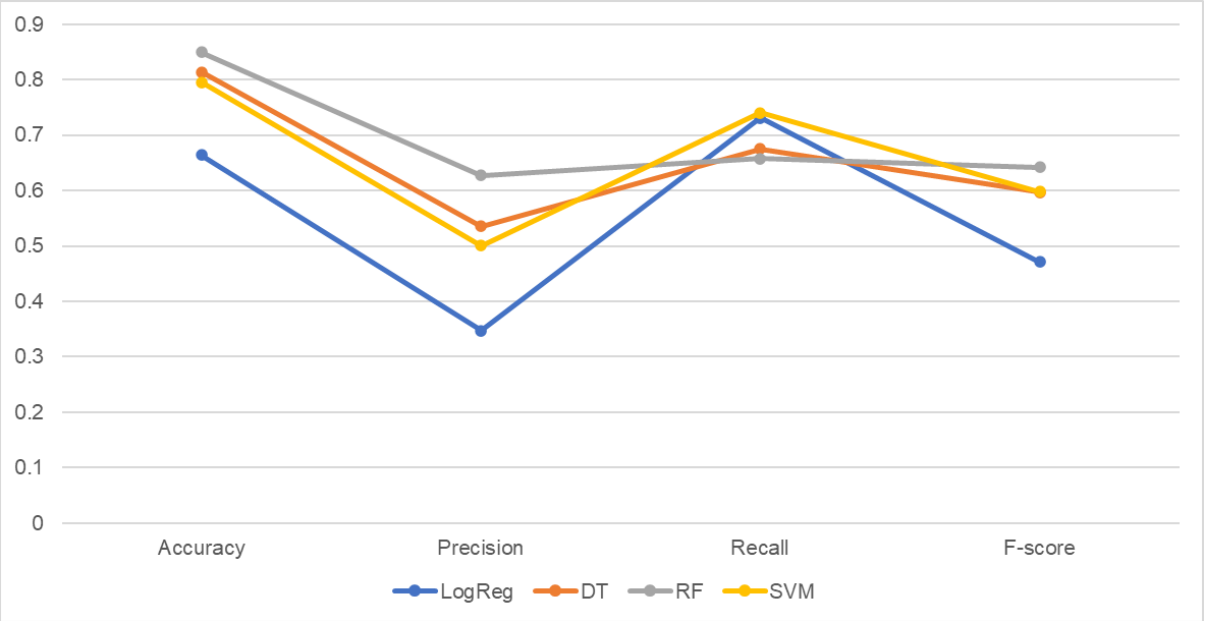

Figure 5. Graphical comparison of performance metrics

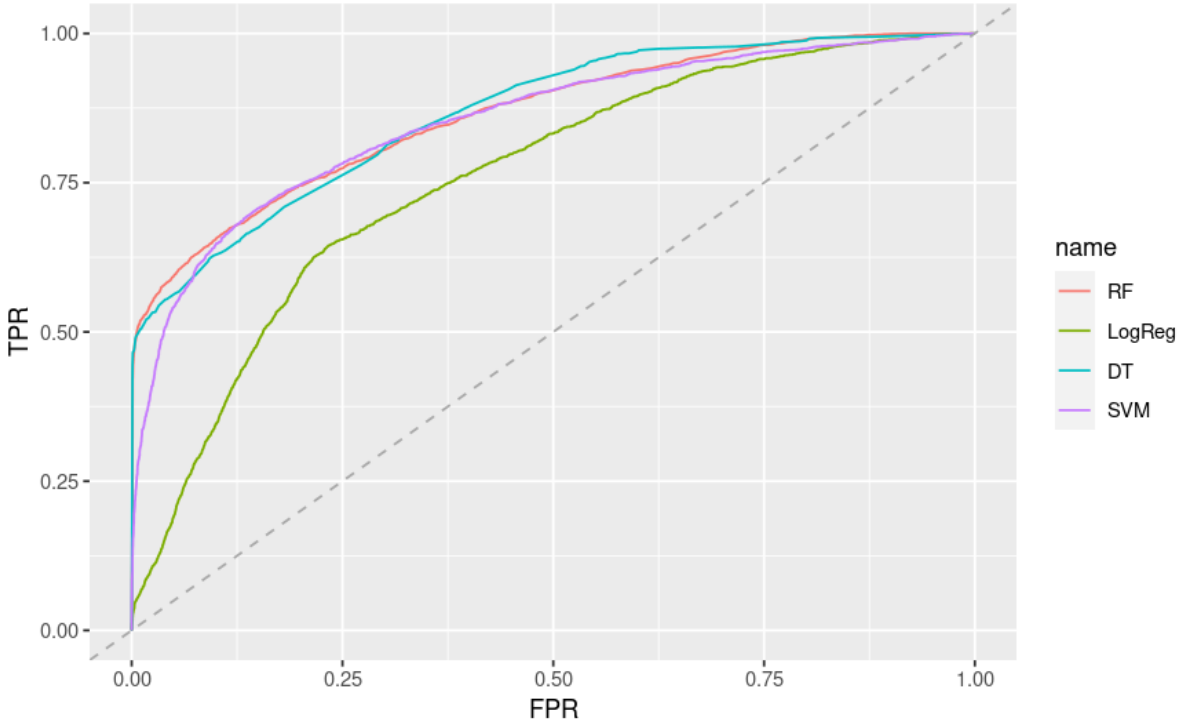

Figure 6. ROC curve compared all models

1914784, 2283394, 2283834, 2286979, 2288495, 2291942

Table 1. AUROC value of all models

	RF	SVM	DT	LogReg
AUC	0.8629	0.8521	0.8651	0.7568

Evaluation of Regressions for Predicting Customer Spend

Both models' performances after training are illustrated in *Table 2* and *Table 3*. LR has a very low R-squared and a very high Residual Standard Error. RF seems to have a slightly better training performance, as the Variance is only 37.5%. However, this is also considered a high variance. The values of these attributes imply that the regressions would predict with a high variance, and neither can be reliable. Therefore, both models are bad to predict and should not be used.

To further check that the models' performance is not desired, the Residual Plot was calculated for each model. If the model is good, the plot should be symmetrically distributed around the 0 line vertically and should not present any patterns. However, the plots for LR (*Figure 7*) and RF (*Figure 8*) are not evenly distributed around the 0 line vertically and have a clear pattern. Therefore, training the model had a bad performance, and it should not be used to predict on the test data.

Table 2. LR performance metrics

	LR
R-squared	0.2689
Residual Standard Error	46.48

Table 3. RF performance metrics

	RF
Mean of squared residuals	1841.279
% Variance	37.5

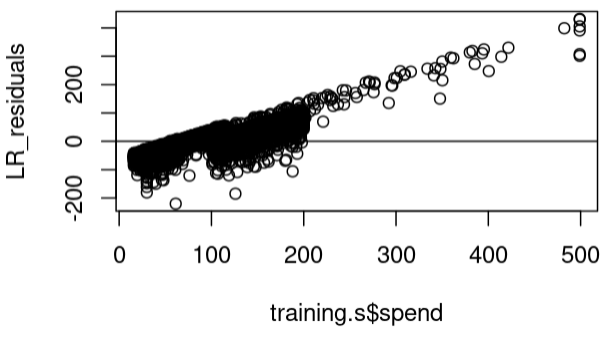

Figure 7. LR Residual Plot

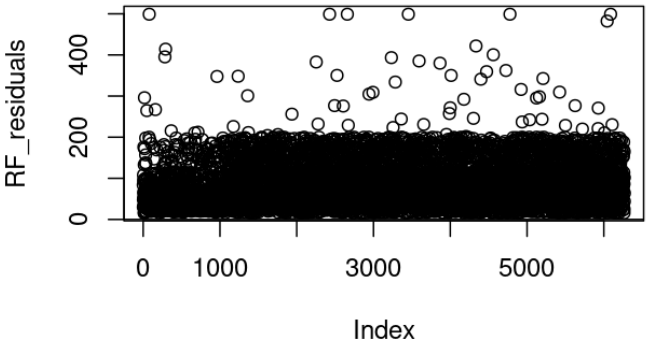

Figure 8. RF Residual Plot

8. Deployment

After training all the data, the predictive model can be used to predict customers' visit rate. With the model, the number of customers visiting the store after receiving an e-mail increased from 15.9% to 62%.

9. Conclusion

In this report, the study used the dataset from 'Universal Plus' of their customers in their previous e-mail marketing campaign in order to predict the right customers for their future strategy so that the company could finally increase the visit rate. After calculating the results using performance metrics and ROC-AUC curve, it was proven that RF modelling techniques gave the best result according to the business requirements. Moreover, the spending prediction was also tried using two regression models. However, the models performed extremely badly, so it was decided not to use them for prediction. Therefore, the additional approach of predicting customer spend was dropped.

For further research, accurate spend prediction models should be added to the existing model to determine both the target group and individual customer importance within the target group based on predicted spending.

1914784, 2283394, 2283834, 2286979, 2288495, 2291942

References

Ali, J. *et al.* (2012) 'Random forests and decision trees', *IJCSI International Journal of Computer Science Issues*, 9(5), pp. 272–278.

Brereton, R. G. and Lloyd, G. R. (2010) 'Support Vector Machines for classification and regression', *Analyst*, 135(2), pp. 230–267. doi: 10.1039/b918972f.

Cui, G. et al. (2008) "Model selection for direct marketing: Performance criteria and validation methods," Marketing Intelligence & Planning, 26(3), pp. 275–292. Available at: https://doi.org/10.1108/02634500810871339.

Gazzah, S., Hechkel, A. and Essoukri Ben Amara, N. (2015). A hybrid sampling method for imbalanced data. *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*. doi:10.1109/ssd.2015.7348093.

Grimes, G. A., Hough, M. G. and Signorella, M. L. (2007) 'Email end users and spam: relations of gender and age group to attitudes and actions', *Computers in Human Behavior*, 23(1), pp. 318–332. doi: 10.1016/j.chb.2004.10.015.

Jain, H., Khunteta, A. and Srivastava, S. (2021) 'Telecom churn prediction and used techniques, datasets and performance measures: a review', Telecommunication Systems, 76(4), pp. 613–630. doi: 10.1007/s11235-020-00727-0.

McCarty, J. A. and Hastak, M. (2007) 'Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression', *Journal of Business Research*, 60(6), pp. 656–662. doi: 10.1016/j.jbusres.2006.06.015.

Nachev, A. and Teodosiev, T. (2015) 'Using Support Vector Machines for Direct Marketing Models', *International Journal of Engineering and Advanced Technology (IJEAT)*, (4), pp. 2249–8958.

Appendix A

Attribute	Definition
Customer_ID	Customer identification number
recency	Months since last purchase before the marketing campaign
purchase_segment	Categorisation for the purchase amount in the past year before the marketing campaign Categories: 1) 0 - 100 : the purchase amount is between 0 and £100 2) 100 - 200: the purchase amount is between £100 and £200 3) 200 - 350; 4) 350 - 500; 5) 500 - 750 6) 750 - 1,000 7) 1,000+
purchase	Actual purchase in the past year before the marketing campaign
mens	Whether the customer purchased men's merchandise in the past year before the marketing campaign (1 = purchased, 0 = not)
womens	Whether the customer purchased women's merchandise in the past year before the marketing campaign (1= purchased, 0 = not)
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 (1 = new customer, 0 = existing customer)
channel	Categorisation of the channels the customer purchased from in the past year. The categories are Phone, Web and Multichannel.
email_segment	E-mail campaign the customer received The categories are 1) Mens E-mail 2) Womens E-mail 3) No E-mail.
age	Age of the customer in years
dependent	Whether the customer has a dependent or not (1 = yes; 0 = no)
account	Whether the customer has an account or not (1 = yes; 0 = no)
employed	Whether the customer has a permanent job (1 = yes; 0 = no)
phone	Whether the customer registered his/her phone or not (1 = yes; 0 = no)

Attribute	Definition
delivery	Categorisation for the delivery address (1 = home; 2 = work; 3 = multiple)
marriage	Marital status (1=married, 2=single, 0 = others)
payment_card	Whether the customer registered a credit card for payment in the past year (1 = yes; 0 = no)
spend	Total amount spent in the following two weeks period
visit	1: the customer visited the shop in the following two weeks period; 0: the customer did not visit the shop in the following two weeks period.