Masters Programmes

Assignment Cover Sheet

Submitted by:	**2257787, 2284896, 2287496, 2288466, 2292957, 2294032**
Date Sent:	**07/12/2022**
Module Title:	**Analytics in Practice**
Module Code:	**IB9BW0**
Date/Year of Module:	**2022**
Submission Deadline:	**08/12/2022**
Word Count:	**2168**
Number of Pages:	**11**

Question: ***E-mail Marketing Project Analytics in Practice***

"I declare that this work is being submitted on behalf of my group, 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 in other assessments for accredited courses of study and if this has been done it may result in us being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work."

Table of Contents

Introduction

Marketing is imperative for the success of any company and to execute a successful marketing campaign it is essential to collect a large sum of customer data. Similarly, all online retailers use data to market their offerings to the target customers to withstand the competitive market and changing trends to increase sales. Since marketing has a cost involved and targeting uninterested customers is a loss for the company, it is crucial to identify potential buyers and contact them to optimize the return on marketing investment.

In order to increase sales, “Universal Plus" has aimed to use direct email marketing to reach out to the buyers. The approach is only productive if the right target audience is engaged, since previously they have had ineffective marketing campaigns due to random selection of customers. The objective of this report is to identify and predict the customers who will visit the shop because of direct email marketing.

Literature Review

To make this report more credible, a few academic papers were reviewed to explore the different data mining techniques and find the most suitable ones.

Everyone knows that “Garbage in, garbage out.”. The main reason for choosing this paper is that data cleaning and preparation is the most important part of the analysis. Big data is influencing business operations and many managers make decisions based on evidence provided by big data analysis. However, most “big data” analysis is problematic due to wrong and careless data cleaning process.

The author (Ridzuan and Wan Zainon, 2019) mentions that there are three important “V”’s in big data:

1.Volume: the definition focuses on the size of the data

2.Variety: refers to the type of data that can be processed which can consist of structured/unstructured data.

3.Veracity: trustworthiness in the data to make a decision (to acquire the right data)

The author also provides several methods for data cleaning and how to deal with mistaken values in the data and insists that the domain knowledge is necessary for cleaning the data and making the analysis more reliable and precise.

Email marketing has been identified as the most preferred form of informing and influencing customers. Predictive marketing can be explained as the development of models based on data analysis, allowing for relevant predictions. Data mining and machine learning have been integrated in digital marketing to personalise marketing offers for each individual customer, leading to better customer retention.
Machine learning is used to analyse information and make intelligent decisions accordingly. Data mining can help marketers recognise patterns and predict customer behaviour from data, using methods such as classification (decision trees and support vector machines (SVM)) and regression. For this study, SVM and decision trees were compared, with the decision tree classifier performing better in all scenarios. Supervised machine learning is used to develop predictive models that are used by businesses to automate email marketing. (Abakouy, En-Naimi and El Haddadi, 2017)

In accordance with consumers adopting new buying behaviours, email marketing is effective for making decisions based on a tailored approach. Alexander A S Gunawan et al. (2016) developed a recommender system with big data and delivered the recommendations through a personalized email as personalization is more powerful than segmentation. An email marketing study by Experian suggests that personalization can increase open rate by 29% and clicks by 41%.
With the focus of personalization, clicks and conversions of target customers are to be predicted by deployment of learning models. The paper is classified based on SVM and Decision Tree machine learning algorithms revealing that the decision tree classifier performs better in all scenarios. As decision trees can be interpreted as a readable form, it aids in decision making. (Abakouy et al., 2019)

Arora et al., 2017 refers to Decision Tree as one of the most applied ‘supervised classification techniques’ that is used in various fields. It represents a method that classifies the categorical data based on many features and it is also used for processing large amounts of data and hence used in data mining applications. This paper highlights the reasons why decision tree is an appropriate method to use. For example, decision trees can be visualized, are simple to understand and interpret, require very little data preparation, and can also work with both

numerical and categorical data. Therefore, this technique is ideal for this project since it consists of both types of data.

SVM is an effective technique and has been used in various domains. This paper refers to the development of credit risk evaluation and bankruptcy prediction models based on linear SVM classifiers. The model performs well in credit risk research and returns high accuracy in predicting the risk of bankruptcy, therefore the use of SVM is ideal for this project.
Even though SVM is applicable in both linear and non-linear problems, the result is not satisfactory since it works best with numerical data. Since most of the elements in this project data are categorical, these factors will affect the model's applicability and effectiveness. Overall, the features of the data are essential when choosing the most appropriate model. The approach of SVM might not be that useful for the data, but it will work well for numerical data. (Danenas and Garsva, 2015)

The authors introduce feature selection techniques to predict customer churn more precisely. It was revealed that in addiction to recency, demographic features and customer information, clickstream/wed logs also act as significant features when discussing about E-retailer. To pinpoint the significant features, this study suggests applying F-ANOVA to calculate their F-score and coefficients from regularized logistic regression to penalize sparse features and to avoid overfitting. The primary purpose of selecting this paper is that there is compelling evidence that these feature selection techniques help to predict customer churn more precisely and are therefore suitable for this project. (Subramanya and Somani, 2017)

Data Preparation:

Data cleaning is one of the most important steps in data analysis. The first step is to check all the NAs in each column revealing that in __purchase segment__ and __spend__ columns have NAs (26 and 49 respectively). In the literature of data cleaning and preparation (Ridzuan and Wan Zainon, 2019), two ways of dealing with NA's are discussed:
1. Fill NAs with suitable values.
2. Remove them.
In the __purchase segment__ column, the values are derived from __purchase__ hence, 'case_when' and 'mutate' should be used to fill those NAs. However, even though the "spend" column was not used in the model, it was important to check whether the observations in the

same rows as the NA values in the spend column were logical, in order to decide whether to keep them in the data. After cleaning the NAs, the target variable __visit__ should be made into a factor variable for future forecasting. Moreover, the use of “duplicated()” is required, to check if there is any duplicated observation in our data.

The next step is variable choosing. Since __Customer ID__ is not a relative variable in the data, it is removed. The __Spend__ column is the consequence of the target variable and hence it is not reasonable to put it in the model. If __Spend__ is put in the forecasting model, the model will assume that every customer who has spent money during last two weeks will visit the store, which is not the case in the real-world. Finally, the company wants a compliment e-mail marketing system to persuade the customers to visit their website, and it is suggested that only the data of the customers who received the campaign should be used. Hence, the customers who did not receive the campaign should be filtered out.

The rest of the variables should be applied for the model training, but to prevent overfitting, the function information.gain() is used and the top ten most informative variables are included in the model. Although the latter version of the model performs worse than the first version, it is assumed that it will be more suitable since it can prevent over-fitting and yield more accurate results when it is applied in the real data set.

Over-sampling, under-sampling and data splitting:

The data is split into training data for model building, and testing data for model evaluation, in a ratio of 3:1.

After the data is split, it is obvious that the population of visit and not visit is heavily imbalanced, therefore over-sampling and under-sampling methods must be applied to the training data.

Nevertheless, the model will perform well when it is trained on data that is similar to the testing data. Based on past data, it is not common for most customers to visit a website only because they receive campaign or promotion messages.

In order to obtain a higher recall (60-70+%), the proportions of 30%, 40%, and 50% were used, however, this resulted in a much lower precision and accuracy rate (lower than 60%).

Thus, it was believed more appropriate to use a proportion of 20%, which is close to the original proportion (15%) and enlarge the training data to 60000 observations so that the model will be exposed to more visit cases to enhance recall and precision rate.

Implementation:

The next step after data preparation, is predicting the target variable. This is done by testing different models.

Extreme Gradient Boost (XGBoost): xgboost() function is used with parameters *max.depth:* Maximum depth of each decision tree, *nthread:* number of CPU threads used, *nrounds*: max number of boosting iterations, *eta:* learning rate of the model.

Support Vector Machine (SVM): svm() function is used with kernel as “radial”, as data is not linearly separated, scale as “TRUE” and probability as “TRUE” to obtain probability class.

Logistic regression: glm() function is used with family as "binomial" due to its binomial distribution.

Decision tree: tree() function is used with argument tree.control and parameter mindev set to zero.

Random Forest: randomForest() function is used with parameters ntree set to 100 and mtry set to 5.

Soft Voting: Voting method based on the probability of each observation predicted by our models.

Model	Accuracy	Precision	Recall	Area Under Curve
XGBoost	87.98%	80.47%	**54.54%**	**86.33%**
RandomForest	87.95%	81.63%	53.13%	82.30%
SVM	86.72%	76.98%	50.21%	81.97%
Logistic	79.40%	40.00%	1.10%	76.71%
Decision Tree	86.42%	84.57%	41.26%	78.53%
Soft Voting	**88.16%**	**88.07%**	48.84%	85.39%

Table 1: Comparison of models

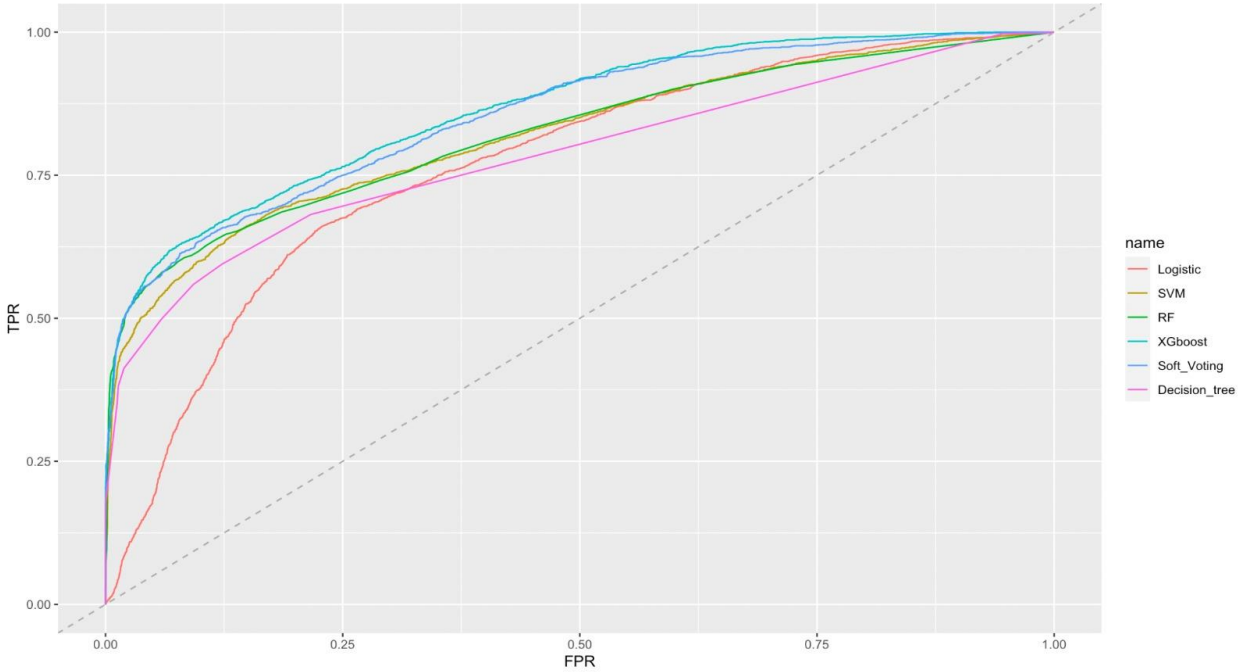

Figure 1: Receiver Operator Characteristic (ROC) Chart

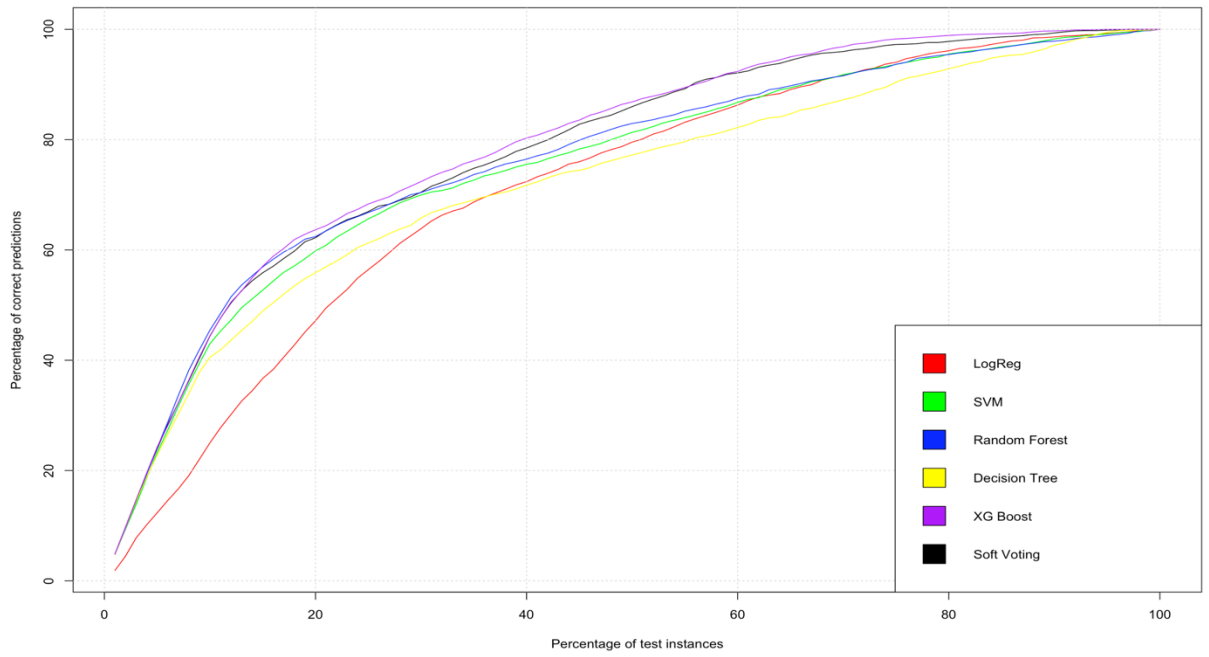

Figure 2: Gain Chart

Evaluation

A perfect classifier is represented by point (0,1) on the ROC graph, meaning that the model can correctly classify with 100% of True Positive and 0% of False Positives. The higher the True Positive Rate (TPR), the better the model performs. The goal is to predict customer churn more precisely.

Furthermore, the area under the graph (AUC) of the ROC graph (Figure 3) is considered to decide which categorization is better. It can be noticed that the AUC of the XGBoost model, 0.863, is the greatest, suggesting that XGBoost is better in comparison to the rest of the models.

Gain is the ratio of cumulative number of positive predictions to the total number of positive predictions in the data. As per Figure 2, XGBoost has the highest gain out of all the models. Additionally, two test instances from Figure 2 were chosen, 20% and 50%, to compare the number of correct predictions made by each model. As per Table 2, XGBoost has the highest number of correct predictions, therefore, it is the best model to use.

Extreme Gradient Boost (XGBoost)

Tree boosting is an effective and widely used machine learning method in data science and XGBoost is a highly scalable end-to-end tree boosting system expected to achieve state-of-the-art results on machine learning challenges (Chen and Guestrin, 2016).
It is a supervised machine learning model that is ideal for classifying and forecasting data. In order to apply XGBoost in R, the data needs to be transformed into a matrix format, which means all characteristic variables should be dummy variables.
The application of decision trees ensembles in XGBoost and tons of parameter tuning, leads to the conclusion that the best fit for the data is 5 tree depths, 0.1 of learning rate, 10 threads, and 100 round.

Conclusion

By using data mining, the company can make a predictive analysis to make the marketing campaign more effective. In this paper, six different mining techniques were used to predict which customers will visit the Universal Plus because of the marketing campaign. The techniques used are XGboost, SVM, Random Forest, Logistic Regression, Decision Tree, and Soft Voting. The results are evaluated, and the performance of these models is compared using the ROC and gain charts, which both suggest that the model which performed best is the XGBoost model. This result is then justified based on accuracy, precision, and recall, the latter being the most important, since it measures the TPR. In conclusion, the recommended model that should be used to predict which customers are more likely to visit the store is the XGBoost model.

References

Abakouy, R., En-Naimi, E.M. and El Haddadi, A. (2017). Classification and Prediction Based Data Mining algorithms to Predict Email Marketing Campaigns. *Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems - ICCWCS'17*. doi:10.1145/3167486.3167520.

Abakouy, R., En-Naimi, E.M., El Haddadi, A. and Lotfi, E. (2019). Data-Driven Marketing: How Machine Learning will improve Decision-Making for Marketers. doi:10.1145/3368756.3369024.

Arora, A., Gupta, B., Rawat, A., Dhami, N. and Jain, A. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining Analysis of Classificat ion Techniques in Dat a Mining. ijesrt journal Dat a Mining Applicat ion in Enrollment Management : A Case St udy Saurabh Pal Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications International Journal of Computer Applications*, 163(8), pp.975–8887.

Chen, T. and Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. [online] Available at: https://arxiv.org/pdf/1603.02754.pdf.

Danenas, P. and Garsva, G. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. *Expert Systems with Applications*, 42(6), pp.3194–3204. doi:10.1016/j.eswa.2014.12.001.

Ridzuan, F. and Wan Zainon, W.M.N. (2019). A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science*, 161, pp.731–738. doi: 10.1016/j.procs.2019.11.177.

Sherazi, S.W.A., Bae, J.-W. and Lee, J.Y. (2021). A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and NSTEMI during 2-year follow-up in patients with acute coronary syndrome. *PLOS ONE*, 16(6), p.e0249338. doi:10.1371/journal.pone.0249338.

Subramanya, K. and Somani, A. (2017). *Enhanced Feature Mining and Classifier Models to Predict Customer Churn for an E-retailer*. pp.531–536.

Appendix

Model <chr>	Percentile <dbl>	cumGainsPercentage <dbl>
Random Forest	20	61.13122
Random forest	50	81.53846
Support Vector Machine	20	59.83569
Support Vector machine	50	81.33272
LogReg	20	47.14742
Logreg	50	79.55272
Decision Tree	20	58.26006
Decision tree	50	79.03748
Extreme Gradient Boost	20	63.57827
Extreme Gradient boost	50	86.80968

Table 2: Gain Comparison

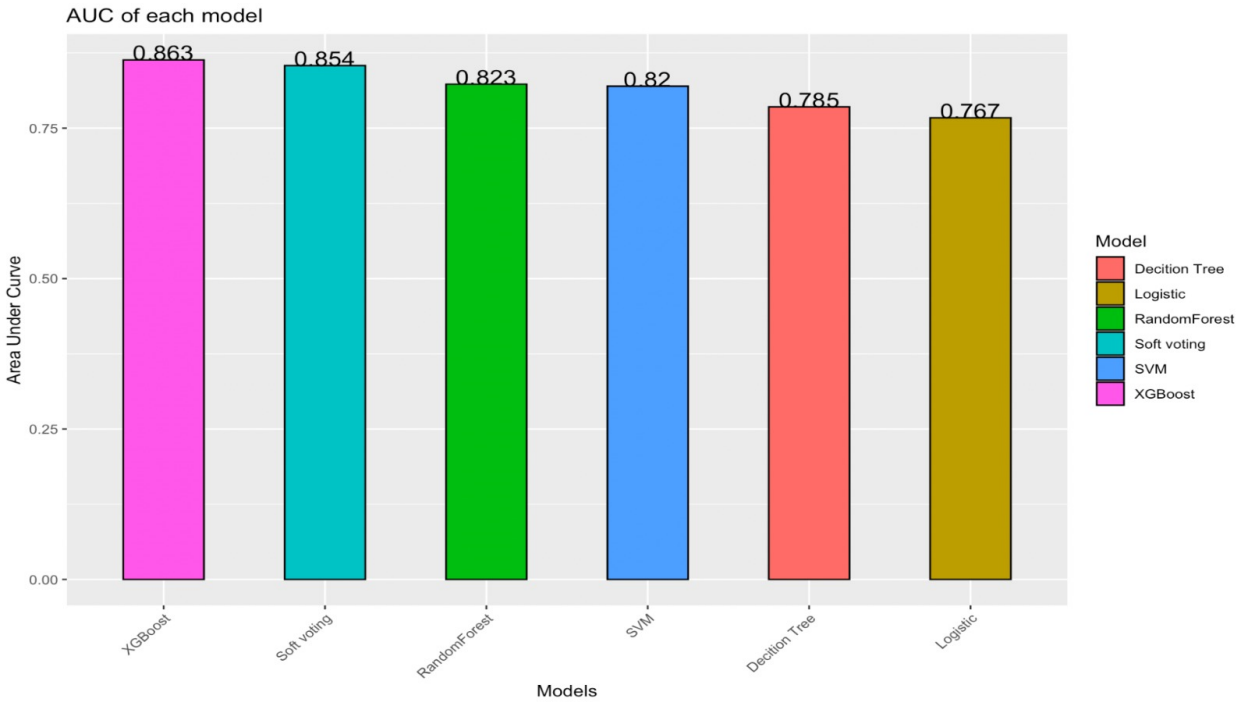

Figure 3: AUC Graph