A Study on a Lead Prediction System for World Plus Using Machine Learning in R

by

Group 7

Advisor: Prof. Nursen Aydin, PHD

Table of content

List of Figures

List of Tables

1. Introduction

World-Plus, a mid-size private bank, requested a proposal to develop a lead prediction system to target prospective customers for their new term deposit product. Figure 1 shows how the CRISP-DM methodology is used in this report. Shi *et al.* (2022) found that classifiers like Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RT) and Logistic Regression (LR) are more effective than statistical methods and will be applied as baseline models in our analysis.

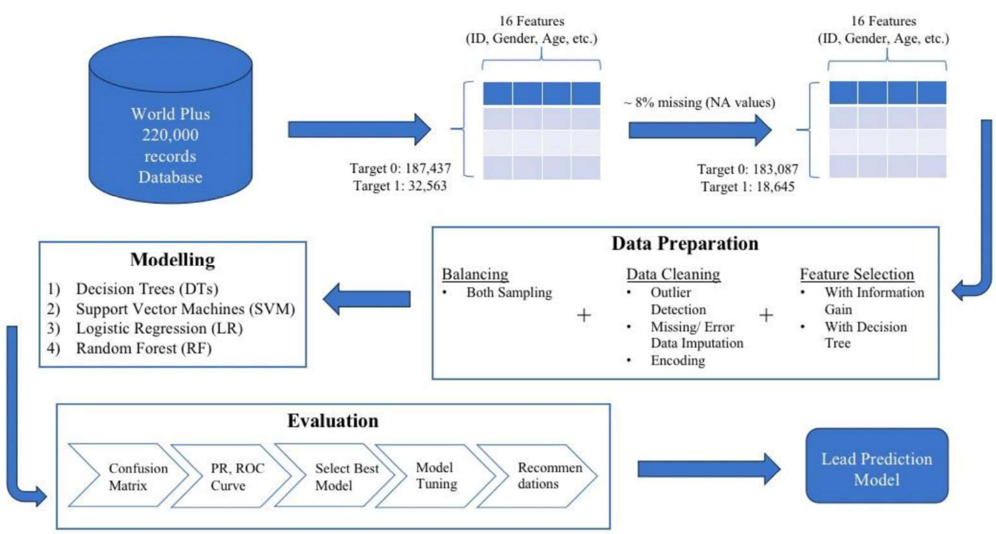

Figure 1 The main steps of this report using CRISP-DM methodology

2. Business Understanding

Studies have found that the banking sector's main challenge is dealing with the overflowing information caused by uncertain speeds and volume in the significant data era (Bedeley, 2014; Hassani *et al.,* 2018). Hence, correctly targeting potential customers for World-Plus will facilitate the efficiency of sales and marketing operations - both in terms of cost and time.

3. Literature Review

Moro *et al.* (2011) proposed a paper regarding the application of CRISP-DM to Bank Direct Marketing, which can be applied to our case (see Appendix A). While some models like DT work well with missing data, others like SVM require missing data deletion or substitution. Therefore, this report tries to manage the instances instead of ignoring the NA value. For data partitioning, it is stated that given the many instances, two out of three of the instances are considered good enough to train the models.

The LR model is a classical classification method under traditional statistical analysis (Tamaddoni *et al.,* 2015). Caigny *et al.* (2022) proposed the logit leaf model (LLM), which is a new hybrid algorithm concept after the improvement of the LR model (see Appendix A). By combining DT feature selection and LR, LLM can retain its advantages and minimize its disadvantages, which provides sufficient evidence for using LR in this report (see Appendix C). Therefore, this report will first select a subset of relevant features through Decision Trees and then use the LR model instead of the linear model to predict.

SVM is a popular prediction method derived from the Vapnik-Chervonenkis (VC) theory, which makes it possible to achieve an optimal classification surface (Shao & Cherkassky, 1999). He *et al.* (2014) chose three models to predict the churn of customers in commercial banks. The result showed that the radial basis function (RBF)

SVM model outperforms the linear SVM model and the LR model, as the kernel function transforms nonlinear classifications into linear by projecting the latitude of samples from low to high (see Appendix C). Similarly, since the customer data provided by World Plus is also personalized and multidimensional (see Appendix A), the RBF SVM model will be chosen for predictive analysis.

RF model can be used for categorical and continuous predictors and response variables (see Appendix C). In the classification data, random forest imputes missing values using the majority value (Adele *et al.,* 2011). This also further supports the argument of using the mode in categorical data to impute missing values. It works on a tree-based ensemble technique and creates a prediction function, which is determined by a loss function and defined to minimize the expected loss value. The essential features are then chosen and used to build the RF model (see Appendix A).

Loss Function Formula - $\sum_{XY}(L(Y, f(X)))$

Song and Lu (2015) mentioned that DT models are one of the best modeling techniques for Data Mining. They have been widely utilized in various fields because they are simple to use and quickly deal with unclean data and missing values. DT models can be used to select the most relevant input features to build later decision trees, which can formulate clinical hypotheses and facilitate subsequent research (see Appendix A and Appendix C). This report will use the DT algorithm to identify critical variables and build one of the models.

For evaluation, Thorleuchter *et al.* (2011)'s paper has been used due to the likeness of our research focus, as it talks about using historic customer data to build models and identify new potential acquisition targets in a business-to-business environment (see Appendix A). The core of this research revolves around finding the optimal approach to estimate the future profitability of these customers by efficiently and effectively targeting them. Given the common objective of developing a successful lead prediction system that accurately identifies leads while avoiding unnecessary costs, attributes and evaluation metrics mentioned in this paper, i.e., Confusion Matrix, Precision-Recall Curve, and Receiver Operator Characteristic (ROC), would be used for the comparative analysis.

In summary, this report combines the listed methods to pinpoint the target customers. Particularly, references are taken from past research that applied the new hybrid algorithm LLM mechanism to predict target customers in the banking industry.

4. Data Understanding

In our research, we obtained a dataset containing 220,000 records of historical customer data with 16 variables, with each variable explained in Table 1. We tried to investigate the relationships between features and the conversion of customers.

Variables	Attribute information
1) ID	customer identification number
2) Gender	gender of the customer
3) Age	age of the customer in years
4) Dependent	whether the customer has a dependent or not
5) Marital_Status	marital state (1=married, 2=single, 0 = others)
6) Region_Code	code of the region for the customer
7) Years_at_Residence	the duration in the current residence (in years)
8) Occupation	occupation type of the customer
9) Channel_Code	acquisition channel code used to reach the customer when they opened their bank account
10) Vintage	the number of months that the customer has been associated with the company.
11) Credit_Product	if the customer has any active credit product (home loan, personal loan, credit card etc.)
12) Avg_Account_Balance	average account balance for the customer in last 12 months
13) Account_Type	account type of the customer with categories Silver, Gold and Platinum
14) Active	if the customer is active in last 3 months
15) Registration	whether the customer has visited the bank for the offered product registration(1 = yes; 0 = no)
16) Target	whether the customer has purchased the product 0: Customer did not purchase the product 1: Customer purchased the product

Table 1 Data Dictionary for World-Plus' Dataset

5. Data Preparation

At first glance, we removed the customer ID column as it does not affect customers' decision to purchase a product. As shown in Table 2, we also checked the data types and encoded them.

5.1 Standardizing the Data

For the "Dependent" variable, some inaccurate data entries with "-1" were found when they were supposed to be "0" or "1". This error constitutes 118 instances, at about 0.05%. It is said that if NA values are less than 1% or up to 5%, they are considered trivial or manageable, while a ratio of over 5% needs to be treated (Elhassan *et al.,* 2021). Thus, we removed all "-1" entries from the "Dependent" column.

Regarding missing values (8% in the "Credit Product" field), we decided to replace them with "No" as it has a significantly higher proportion than "Yes." This treatment was based on Silva-Ramírez *et al.*'s paper (2010), where "qualitative variables like NA can be imputed with the mode." Table 2 shows the list of selected variables.

Feature Name	Data Type / Values	Feature Name	Data Type / Values
Gender	Factor : "0","1"	Vintage	Integer : 38 49 88 ...
Age	Integer : 73 30 56 ...	Credit_Product	Factor : "0","1","2"
Dependent	Factor : "0","1"	Avg_Account_Balance	Integer : 1045696 581988 ...
Marital_Status	Factor : "0","1","2"	Account_Type	Factor : "1","2","3","4"
Region_Code	Factor : "RG250","RG251", ...	Active	Factor : "0","1","2"
Years_at_Residence	Integer : 1 3 5 ...	Registration	Factor : "0","1","2"
Occupation	Factor : "1","2","3","4"	Target	Factor : "0","1","2"
Channel_Code	Factor : "1","2","3","4"		

Table 2 Structure of Variables by Data Type and Values

5.2 Data partitioning and balancing

For the partitioning part, we divided the datasets into training and test by the proportion of ⅔ and ⅓, respectively, with stratified sampling, which was considered good enough to build the models (Moro *et al.,* 2011). Since the dataset was skewed, we decided to use the both-sampling method. Seiffert *et al.* (2008) explained that this hybrid method usually improves performance compared to using only a single sampling procedure, as fewer observations are removed from the data, decreasing the loss of information.

5.3 Features selection

Next, we made feature selection through a two-step process, as it is an essential phase in pattern recognition and model performance (Zhou *et al.,* 2020; V. *et al.,* 2006; Sadhasivam *et al.,* 2021). We first used the information gain function with the median method and then applied the DT model for further selection. Consequently, the top four features are Age, Registration, Vintage, and Channel Code, as they improve understanding of customer preferences, customer behavior, temporal dimension, and guiding force (Stanley *et al.,* 1985; Tao & Rosa Yeh, 2003; Robertson et al., 1998; Goić, Jerath, & Kalyanam, 2022). Here, only the top four most dominant features are identified as suggested by the J48 algorithm (V. *et al.,* 2006) (see Appendix D). Since classification is the process of labeling and categorizing the input data, it is assumed that the provided data categories can predict the target variable using the models above. Thus, we hold the assumption that the chosen variables are interconnected.

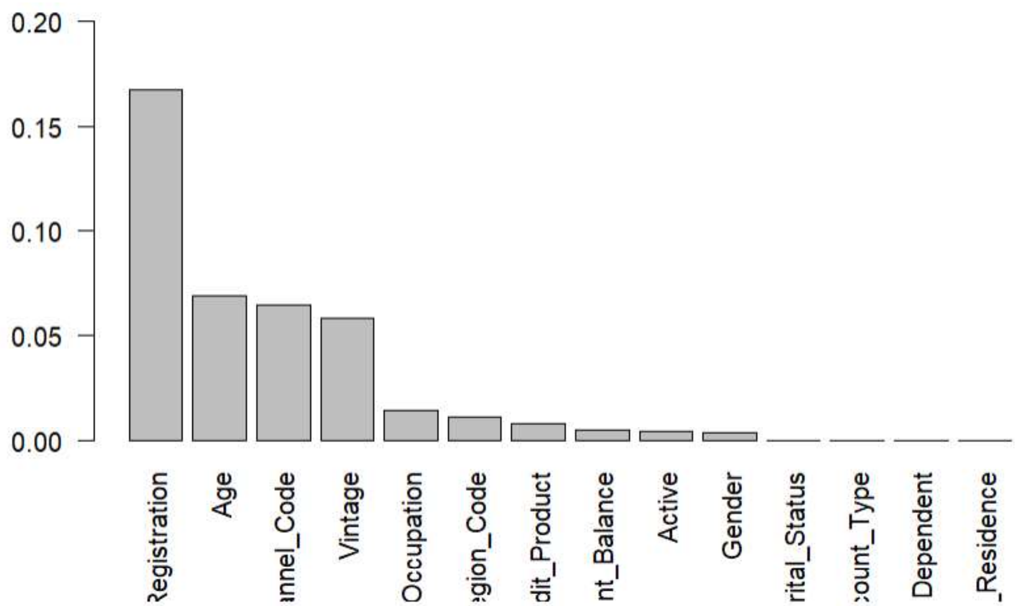

Figure 2 Information gain values of each variable

6. Modelling

By filtering the features, a feature subset can be generated and used in the four selected classifiers, including the LR, RBF SVM, DT, and RF models. Due to respective advantages and limitations, each of the four models outperformed in different scenarios (see Appendix C).

The RBF SVM model can correctly separate multiple dimensions and maximize their boundaries (Xia & Jin, 2008; Shabankareh *et al.,* 2021). It becomes beneficial to use the C5.0 algorithm because it can perform feature selection and give high accuracy with low memory usage (Pandya *et al.,* 2015). For RF, the report will consider a higher number of trees to ensure less out-of-bag error and avoid data pruning (Adele *et al.,* 2011).

7. Results and Evaluation

The model-evaluation process involves identifying potential candidates and model-tuning to find the best-fitting model.

7.1 Confusion Matrix

The confusion matrix can calculate several evaluation metrics (See Appendix B). Table 3 shows the results of 5 key attributes.

Model	**Instances**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Error Rate**
Logistic Regression	65964	0.8911	**0.6571**	0.5495	0.5985	0.1089
SVM	65964	**0.8931**	0.6518	0.5935	**0.6212**	**0.1069**
Decision Tree	65964	0.8538	0.5041	**0.6371**	0.5629	0.1462
Random Forest	65964	0.8675	0.5460	0.6121	0.5772	0.1325

Table 3 Comparison of models between classification metrics before model tuning

In the case of an imbalanced dataset, classifiers are often more inclined to predict the majority class correctly. Hence, general classification rules like accuracy fail to measure the model's predictive power effectively (He *et al.,* 2014). As such, more emphasis will be placed on precision and recall.

A more considerable precision indicates that the model can correctly classify the observations and, in this case, cut costs of uninterested customers.

$$Precision = \frac{TP}{TP + FP}$$

A higher recall indicates a higher proportion of accurate labels being identified and not missing any potential customers.

$$Recall = \frac{TP}{TP + FN}$$

When precision and recall are similar across different models and no single model outperforms the rest, the F1 Score is used to find the right balance between the two.

$$F1\ Score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

From Table 3, SVM has the highest F1 score of 0.6212, followed by LR (0.5985), RF (0.5772) and DT (0.5629). Given World Plus's objective of targeting prospective customers and avoiding unnecessary expenses on uninterested customers, a balanced model with the highest F1 score is preferred.

To further explore precision and recall, the error rate will be examined (Das, 2015).

$$Error\ Rate\ = \frac{FP + FN}{TP + TN +\ FP + FN}$$

Once again, SVM has the lowest error rate of 0.1069, followed by LR (0.1089), RF (0.1325) and DT (0.1462).

7.2 Receiver Operating Characteristic Curve (ROC) and area under (AUC)

The ROC curve (See Figure 2) is a visual representation depicting the effectiveness of binary classifiers by comparing the true positive rate (TPR) to the false positive rate (FPR); as stated by Zhang *et al.* (2015), it is a valuable instrument in evaluating paired classifiers.

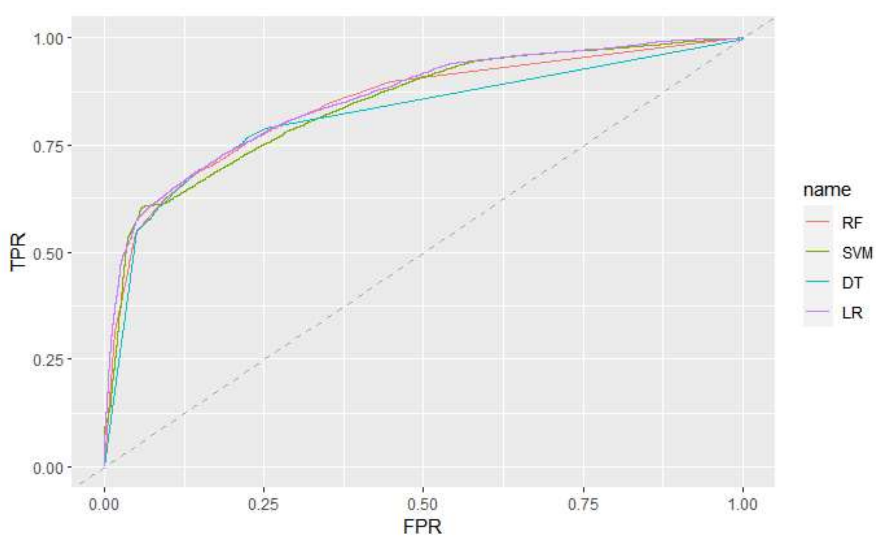

Figure 3 ROC curves of all models

As shown in Figure 3, we compared the ROC curves for all models. However, according to Drummond & Holte (2000), ROC curves have significant disadvantages when evaluating imbalanced data models, as the curve does not explicitly represent decision thresholds.

In this context, ROC curves may not provide a complete picture of classifier performance because the TPR is calculated based on the number of positive instances and the majority class, not the minority class (Manel *et al.,* 2001).

Models	**Area under ROC curve (AUC)**
Decision Tree	0.8192
Random Forest	0.8435
Logistic Regression	**0.8557**
SVM	0.8434

Table 4 Area under ROC curve (AUC) across respective models

From table 4, LR has the highest AUC (0.8557), followed by RF (0.8435), SVM (0.8434) and DT (0.8192).

7.3 Precision recall (PR) curve and area under (AUCPR)

Models	**Area under PR curve (AUCPR)**
Decision Tree	0.5151
Random Forest	0.5683
Logistic Regression	0.6265
SVM	0.5855

Table 5 Area under PR curve (AUCPR) across respective models

The PR curve is a better alternative to the ROC curve, which highlights performance differences lost in ROC curves (Goodrich *et al.,* 2006). It must incorporate correctly predicted instances and be more prone to exaggerate model performance for unbalanced datasets (Sofaer *et al.,* 2018).

Although LR has a larger AUCPR, the optimal points for SVM and LR are almost identical. For instance, if a 0.605 recall benchmark was chosen, LR and SVM will have a precision of about 0.646 (See Figure 6 & 7). As for DT and RF, the precision is just about 0.5 (See Figure 4 & 5). The SVM model is still chosen since we value the equal balance and performance between both elements.

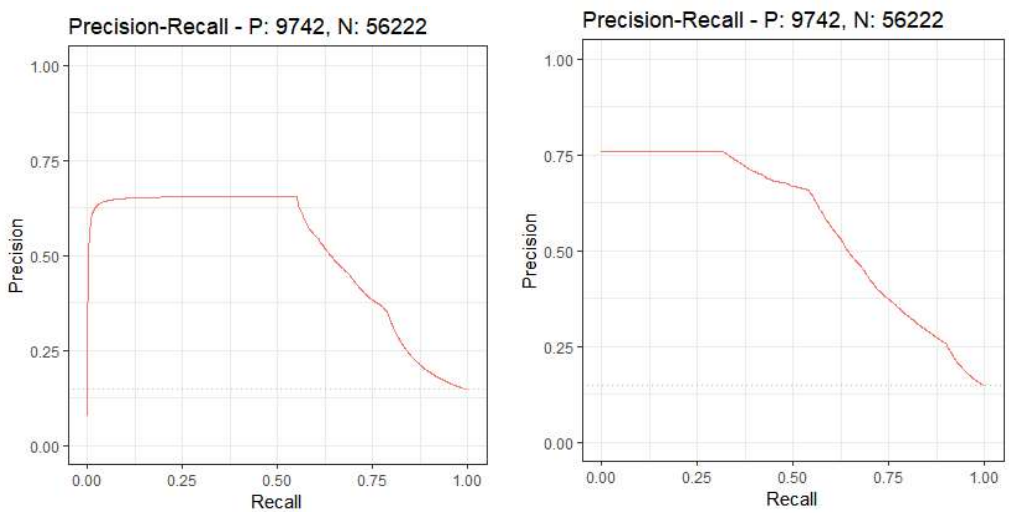

Figure 4 & 5 Precision-Recall Curves for Decision Tree (left) and Random Forest (right)

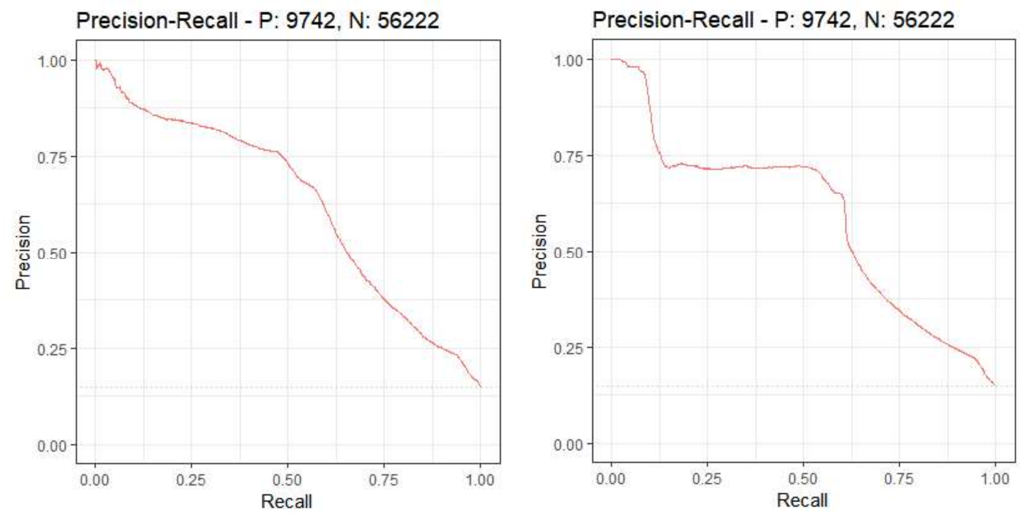

Figure 6 & 7 Precision-Recall Curves for Logistic Regression (left) and SVM (right)

Summarizing the results, SVM outperforms the others by having the highest F1 score and lowest error rate. Precision and recall are not the greatest, but a well-balanced model is the first prioritization.

8. Conclusion

This report applied CRISP-DM methodology to improve the lead prediction system. The dominant four features we suggested for the bank are Age, Registration, Vintage, and Channel Code. As World Plus embarks on lead prediction, these dominant features and the SVM model synergy become a powerful guide, steering strategic decisions and optimizing outcomes for sustained success. For further improvement, model tuning is suggested.

9. References

1. Bedeley, R. (2014) *Big Data Opportunities and challenges: The case of banking industry, AIS Electronic Library (AISeL)*. Available at: https://aisel.aisnet.org/sais 2014/2/ (Accessed: 20 November 2023).

2. Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble Machine Learning*, 157–175. doi:10.1007/978-1-4419-9326-7_5

3. Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. Ensemble Machine Learning, 157–175. doi:10.1007/978-1-4419-9326-7_5 De Caigny, A., Coussement, K. and De Bock, K.W. (2018) 'A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees', *European Journal of Operational Research*, 269(2), pp. 760–772. doi:10.1016/j.ejor.2018.02.009.

4. Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/347090.347126

5. Elhassan, A. et al. (2021) *ILA4: Overcoming missing values in machine learning datasets – an inductive learning approach, ScienceDirect.* Available at: https://www.sciencedirect.com/science/article/pii/S1319157821000501 (Accessed: 20 November 2023).

6. Goadrich, M., Oliphant, L., Shavlik, J. (2006) 'Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves', *Machine Learning,* 64, pp. 231–262

7. Goić, M., Jerath, K., & Kalyanam, K. (2022). The roles of multiple channels in predicting website visits and purchases: Engagers versus closers. *International*

Journal of Research in Marketing, *39*(3), 656–677. doi:10.1016/j.ijresmar.2021.12.004

8. Hassani, H., Huang, X. and Silva, E. (2018) *Digitalisation and big data mining in banking, MDPI.* Available at: https://www.mdpi.com/2504-2289/2/3/18 (Accessed: 20 November 2023).
9. He, B. *et al.* (2014) 'Prediction of customer attrition of commercial banks based on SVM model', *Procedia Computer Science*, 31, pp. 423–430. doi:10.1016/j.procs.2014.05.286.
10. Manel, S., Williams, H. C., & Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: The need to account for prevalence. *Journal of Applied Ecology*, 38, 921–931.
11. Moro, S., Laureano, R. and Cortez, P. (2011) *Using data mining for Bank Direct Marketing: An application of the CRISP-DM methodology*, *RepositoriUM.* Available at: https://repositorium.uminho.pt/handle/1822/14838 (Accessed: 07 November 2023).
12. Pandya, R. (2015) *C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. International Journal of Computer Applications,* 117(16). Available at https://www.academia.edu/download/57522627/15dbc3f753304c8ea8ebebc677b3b6b12125_1.pdf SONG, Y.-y., 2015.
13. Sadhasivam, J. et al. (2021) *Diabetes disease prediction using decision tree for feature selection, IOP Science.* Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1964/6/062116/meta?gclid=CjwKCAiAvJarBhA1EiwAGgZl0PVaI8VNI8S-I7xSRM93FQ9Iwj-4ViMx4W-I-JN5y99ZJQgQD78cexoCNqoQAvD_BwE (Accessed: 28 November 2023).

14. Seiffert, C., Khoshgoftaar, T.M. and Hulse, J.V. (2008) *Hybrid sampling for imbalanced data, IEEE Xplore.* Available at: https://ieeexplore.ieee.org/document/4583030 (Accessed: 20 November 2023).
15. Shabankareh, M.J. *et al.* (2021) 'A stacking-based data mining solution to customer churn prediction', *Journal of Relationship Marketing*, 21(2), pp. 124–147. doi:10.1080/15332667.2021.1889743.
16. Shao, X. and Cherkassky, V. (1999) 'Multi-resolution support Vector Machine', IJCNN'99. *International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339) [Preprint].* doi:10.1109/ijcnn.1999.831103.
17. Shi, S., Tse, R., Luo, W., D'Addona, S., Pau, G.: Machine learning-driven credit risk: a systemic review. *Neural Comput.* Appl. 34, 14327–14339 (2022))
18. Silva-Ramírez a, E.-L. et al. (2010) *Missing value imputation on missing completely at random data using multilayer perceptrons, ScienceDirect*. Available at: https://www.sciencedirect.com/science/article/pii/S0893608010001735?fr=RR-2&ref=pdf_download&rr=82d2f36bda317786 (Accessed: 19 November 2023).
19. Sofaer R. Helen, Hoeting A. Jennifer, Jarnevich S. Catherine S. (2018). The area under the precision-recall curve as a performance metric for rare binary events. *British Ecological Society,* pp. 565. https://doi.org/10.1111/2041-210X.13140
20. Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130.
21. Stanley, T. O., Ford, J. K., & Richards, S. K. (1985). Segmentation of bank customers by age. *International Journal of Bank Marketing*, *3*(3), 56–63. doi:10.1108/eb010761
22. Sugumaran, V., Muralidharan, V. and Ramachandran, K.I. (2007) *Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry,* 27(2), p. 130. available at : https://doi.org/10.11919%2Fj.issn.1002-

0829.215044" \t "_blank

23. T. K. Das. (2015) "A customer classification prediction model based on machine learning techniques," *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pp. 321-326, doi: 10.1109/ICATCCT.2015.7456903.
24. Tamaddoni, A., Stakhovych, S. and Ewing, M. (2015) ‘Comparing churn prediction techniques and assessing their performance’, *Journal of Service Research*, 19(2), pp. 123–141. doi:10.1177/1094670515616376.
25. Tao, Y.-H., & Rosa Yeh, C.-C. (2003). Simple database marketing tools in customer analysis and Retention. *International Journal of Information Management*, *23*(4), 291–301. doi:10.1016/s0268-4012(03)00052-5
26. Thorleuchter, Dirk, Van den Poel, Dirk, Prinzie, Anita. (2011). Analyzing existing customers’ websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Elsevier Ltd.* https://doi.org/10.1016/j.eswa.2011.08.115
27. V., S., V., M. and K.I., R. (2006) *Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of Roller Bearing, ScienceDirect.* Available at: https://www.sciencedirect.com/science/article/pii/S0888327006001142 (Accessed: 28 November 2023).
28. XIA, G. and JIN, W. (2008) ‘Model of customer churn prediction on support vector machine’, *Systems Engineering - Theory & Practice*, 28(1), pp. 71–77. doi:10.1016/s1874-8651(09)60003-x.
29. Xiahou, X. and Harada, Y. (2022) ‘B2C e-commerce customer churn prediction based on K-means and SVM’, *Journal of Theoretical and Applied Electronic Commerce Research*, 17(2), pp. 458–475. doi:10.3390/jtaer17020024.

30. Zhang, X., Li, X., Feng, Y. & Liu, Z. (2015), ‘The use of roc and auc in the validation of objective image fusion evaluation metrics’, *Signal processing* 115, 38–48.
31. Zhou, H. et al. (2020) *A feature selection algorithm of decision tree based on feature weight, ScienceDirect.* Available at: https://www.sciencedirect.com/science/article/pii/S0957417420306515 (Accessed: 28 November 2023).

10. Appendix

Appendix A. The supplement to the literature review and corresponding justifications

Authors	Title	Year	What?	Techniques	Justifications
Moro, S., Laureano, R. and Cortez, P.	Using data mining for Bank Direct Marketing: An application of the CRISP-DM methodology	2011	This paper implemented DM project based on the CRISP-DM methodology. The data were collected from a Portuguese marketing campaign related with bank deposit subscription. The objective is to find a model that can develop campaign efficiency by identifying the main characteristics that help classify potential buying customers.	CRISP-DM, Three DM algorithms (i.e. NB, DT and SVM), AUC plot, ROC analysis	• The paper includes all the processes from data preparation to model evaluation; this allows us to understand more about how CRISP-DM can be used. • During the Data Preparation phase of the paper, there were also lots of observations with missing values that were dealt with; it was explained that while some models like Decision Trees work well with missing data, there are others like SVM that require missing data deletion or substitution. As a result, instead of ignoring the NA value, we try to find a way to deal with the instances that contain missing values. • For the part of data partitioning the article explained that as there are many instances, two out of three of the instances are considered good enough to build the

					models. Therefore, the authors and our group randomly divided the datasets into training and test by the proportion of ⅔ and ⅓ respectively.
Benlan He,Yong Shi, Qian Wan , and Xi Zhao	Prediction of customer attrition of commercial banks based on SVM model	2014	Comparation between linear SVM model, radial basis function (RBF) SVM and logistic regression model on predicting the churn of customers in commercial bank	Support vector machine (SVM), Logistic regression (LR)	• The paper and this report have similar data qualities, such as high dimension and personalization. • The advantages and disadvantages of different SVM types and logistic regression and the model performance situation contribute to the model selection in this paper.
Arno De Caigny, Kristof Cousseme nt and Koen W. De Bock	A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees	2018	This paper introduced a new hybrid algorithm. Logit leaf model contains two stages: a segmentation stage dominated by decision tree and a prediction stage dominated by logistic regression.	Logistic regression (LR), Random forests (RF), Logistic model trees (LMT), Decision tree (DT) Logistic Leaf model (LLM)	• The analysis logic of LLM can reduce data heterogeneity. • Through the feature selection of DT, the LR model can analyze the corresponding subsets to reduce the interactions between variables. • As one of the innovations of this paper, the feature selection led by the leading DT model is more convincing than the common information gain method. • The target variable is binary (refers to 1 and 0).

					This report cannot choose the linear model because it will make the target variable fall outside of the range.
Dirk Thorleuchter, Dirk Van den Poel, Anita Prinzie	Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing	2011	Investigates the issue of predicting new customers as profitable based on information about existing customers in a business-to-business environment.	Evaluation metrics: precision, recall, area under the receiver operating characteristics curve (AUC), sensitivity, and specificity	• The paper focuses on using predictive analytics to help identify new potential acquisition targets. • Although the authors conducted analysis on textual information of existing customers' websites, the ultimate goal of both predictive systems are highly similar, hence useful evaluation metrics from this paper will be extracted..
Yan-yan SONG ,Ying LU	Decision tree methods: applications for classification and prediction	2015	Advantage of using Decision tree algorithm with feature selection.	CART, C4.5, CHAID, and QUEST	• This paper describes how the DT model works well with missing data and also provides important features which makes the model less complex . • Pruning method is used to find the optimal size of DT if the dataset is very large with lots of variables • Stopping rules must be applied to the DT model , to avoid overfitting .

Rutvija Pandya	C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning	2015	In this research work the framework proposed used C5.0 classifier that Performs highlight determination and diminished mistake pruning methods which are depicted in this paper.	In this research work, comparison between ID3 C4.5 and C5.0 is presented.	• Decision trees can handle both numerical and categorical data without extensive data processing. • The tree structure presents a series of direct choices, making it an important algorithm for understanding and visualizing. • While analyzing different model results, it was easier to get an idea about the attribute weightage of different variables by looking at the tree model.
Adele Cutler, David Richard Cutler, John R Stevens	Random Forest	2011	This paper covers the algorithm and how it performs differently in classification problem, variable importance, missing value imputation, out of Bag data and tuning hyperparameters of model	Random Forest, Confusion Matrix, Tuning	• The paper justifies that Random Forest is appealing as it measures variable importance, class weightage and can also treat missing values. • The paper explains that loss function is a measure of how close is f(X) to Y and works on a zero – one loss model for classification. The esemble constructs f in terms of collection of “base learners” which are combined to give predictors. Y which is the response variable is the most frequently predicted class f(x) in classification.

					• The paper emphasis on a common misconception in case of calculating Out – of – Bag error rate in similar problem as classification which is computing by averaging error rates for each tree. Instead, we can use error rate of out – of – bag predictions. This helps us to obtain class wise error rate for each class. • The paper explain through a graph inverse relationship between number of trees and out of bag error rate. We have considered 500 no. of tree to create a model which is computationally efficient with minimum error rate. • Tuning (to update basis if we will tune RF model or not)

Appendix B. Confusion Matrices to Respective Models

		Actual	
		0	**1**
Predicted	**0**	53133 (TN)	3960 (FN)
	1	3089 (FP)	5782 (TP)

Table 1. Confusion Matrix for SVM

		Actual	
		0	**1**
Predicted	**0**	51264 (TN)	3779 (FN)
	1	4958 (FP)	5963 (TP)

Table 2. Confusion Matrix for Random Forest

		Actual	
		0	**1**
Predicted	**0**	50116 (TN)	3535 (FN)
	1	6106 (FP)	6207 (TP)

Table 3. Confusion Matrix for Decision Tree

		Actual	
		0	**1**
Predicted	**0**	53429 (TN)	4389 (FN)
	1	2793 (FP)	5333 (TP)

Table 4. Confusion Matrix for Logistic Regression

	TPR	TNR	FPR	FNR
SVM	0.5485	0.9503	0.0497	0.4515
Decision Tree	0.6371	0.8914	0.1086	0.3629
Random Forest	0.6121	0.9118	0.0882	0.3879
Logistic Regression	0.5935	0.9451	0.0549	0.4065

Table 5. True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative Nate (FNR) for all models

$$TPR = \frac{TP}{TP + FN} \qquad TNR = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{FP + TN} \qquad FNR = \frac{FN}{FN + TP}$$

Appendix C. The Supplements of Techniques

Techniques	**Definitions**	**Advantages**	**Disadvantages/ limitations**	**Formulars**	**References**
Decision Tree	The use of decision tree methodology is prevalent in data mining, where it is commonly utilized to construct classification systems by considering multiple covariates or to develop prediction algorithms for a specific target variable	Easy to understand and present. Able to handle missing values.	DT mode can be subject to overfitting and underfitting, while working with small dataset. Strong correlation between different input variables can lead to inaccurate presentation of the results		SONG, 2015
Support Vector Machine	SVM model is a machine learning based prediction method derived from the Vapnik-Chervonenkis (VC) theory, which makes it possible to achieve an optimal classification surface. By finding a hyperplane that satisfies the classification requirements, this popular machine learning method has the advantage of being able to correctly separate multiple	Optimize nonlinear decision boundaries via the kernel function, which improves generalization and avoids overfitting. They argued that the kernel function can transform nonlinear classification into linear classification by projecting the latitude of samples from low to high.	Typically, compared to DT and RF, SVM model perform badly when dealing with multidimensional data because of the inability to perform feature filtering and combination processing.	$L(w,b,a) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} a_i\left(1 - y_i\left(w^T x_i + b\right)\right)$ $W = \sum_{i=1}^{m} a_i x_i y_i = 0$ $\sum_{i=1}^{m} a_i y_i = 0$ [illegible] $i = 1, 2, \ldots, m$	Shao and Cherkassky, 1999; XIA and JIN, 2008; Shabankareh et al., 2021

	dimensions and maximize their boundaries.				
Logistic regression	LR model is a classical classification method under traditional statistical analysis. It can predict the probability of an unknown category in the data by combining the categories already present in the data.	Solve and apply to problems related to continuous and categorical variables.	LR model cannot recognize and handle interactions between variables.	$P(Y=1\|X)=\frac{exp(wx+b)}{1+exp(wx+b)}$ $P(Y=0\|X)=\frac{1}{1+exp(wx+b)}$ $P(Y=1\|X)=\frac{exp(wx+b)}{1+exp(wx+b)}$ $P($	Xiahou and Harada, 2022; Tamaddoni, Stakhovych and Ewing, 2015
Random Forest	Random forest is a tree – based ensemble in which each tree depends on collection of random variable. The combination of variables are used to get the response.	It can measure importance of each feature for the training data. It can handle both classification and regression. It depends on only 2-3 tuning parameters. Random components are based on 2 main factors – number of trees using bootstrap sample from original data and other is splitting of variables for each tree randomly	It can be biased in favor of attributes with different number of levels. Pruning might not work best to overcome overfitting in Random forest.	D = {(x1, y1), . . . , (xN, yN)} xi = (xi,1, . . . , xi,p)T *P = all variable predictors* *k = terminal node* ˆh(x) = *arg*maxyå*ni* =1 *I*(*yki* = *y*) for classification, where *I*(*yki* = *y*) = 1 if *yki* = *y* and 0	Adele , John & David, 2011

Appendix D. Features selection techniques (a two-step process)

Step	**Justifications**
Step 1: Information gain function with the median method	We used the information gain function with the median method to filter out seven features or half of all features since the DT model is sensitive to irrelevant features which leads to less classification accuracy. We chose the median in feature weight as the threshold since setting the appropriate threshold value needs experience and experiments; using the adaptive method could help divide the features into two equal parts, high and low correlation (Zhou *et al*., 2020).
Step 2: Decision Trees model	Sugumaran *et al*. (2007) argued that the features which do not contribute significantly can be removed by deciding on a suitable threshold. Reducing the unwanted features also reduces the complexity of the model. Therefore, we applied the DT model for further selection as the model works based on the information gain of the features; it is said that only those contributing to the classification appear. The result of attribution usage from DT is below: Attribute usage: 100.00% Age 100.00% Registration 96.01% Vintage 70.37% Channel_Code 52.24% Occupation 49.27% Region_Code 30.57% Credit_Product