Table of Contents

1. Introduction

1.1 Business background

World Plus, a mid-sized private banking institution, confronts the challenge of precision in lead identification. The conversion of leads is a crucial factor in the success of banks, requiring the adept identification and targeting of potential customers with the highest likelihood of conversion.

World Plus extends a diverse portfolio of banking products, including loans, investment instruments, savings accounts, and credit offerings. Currently, World Plus employs various communication channels to effectively promote its financial products to customers. These channels include call centres, live chat, email, and social media. By employing the omnichannel approach, World Plus ensures that its customers receive tailored and timely information about the diverse range of banking products.

1.2 Objective

The objective of this project is to develop an accurate lead prediction system to target prospective customers for World Plus's new term deposit product by following the CRISP-DM methodology, working through business understanding, data understanding, preparation, modelling, evaluation, and deployment, empowering the bank to reduce its marketing and sales costs and increase its lead conversion rate.

2. Literature Review

2.1 Handling Missing Data

Zhuchkova and Rotmistrov (2022) explained how complete case analysis (CCA) is used for Missing Completely At Random (MCAR), Missing Not At Random (MNAR), and Multiple Imputation (MI) for Missing at Random (MAR). Missing-Indicator Method (MIM) could be used strategically for MNAR and MCAR, mainly when adding a new variable brings forth the reasons behind missing data. These insights help us address the missing value issue.

2.2 Dealing with Class Imbalance Problem

Wongvorachan, He, and Bulut (2023) compared the performance of under-sampling, over-sampling, and hybrid methodologies for working on imbalanced classification. The study shows that for moderately imbalanced data, over-sampling is useful. Hybrid methods are the most useful for highly imbalanced data sets. This information helps us in using the correct approach for dealing with imbalanced target variables.

2.3 Selecting Significant Variables

Rongjie and Mohamed (2013) employed the Support Vector Machine (SVM) to evaluate real-time crash risk in Active Traffic Management. The Classification and Regression Trees (CART) model has been developed to select crucial explanatory variables by calculating the frequency each variable appeared and its relative position in Decision Trees (DT). Since the SVM model lacks the capability of selecting significant variables, using CART for feature selection before SVM yield better model performance; this knowledge is employed in this project.

2.4 Data Mining Model Selection and Optimisation

Moro et al. (2014) study on bank telemarketing success prediction compared the performance of Logistic Regression (LR), DT, Neural Network (NN), and SVM, concluding that NN performed the best. Apampa (2016) proposed the application of Random Forest (RF) on bank datasets with multiple categorical variables and emphasised that balancing data sets improve the prediction results. This information is used to help select the Data Mining (DM) tools used in this project.

2.5 Random Forest as an Alternative to Regression

He et al. (2018) compared the overall performance of LR and RF. They proposed that RF is a superior method for prediction because of its easy application, lower computational cost, higher predictive accuracy, and better interpretability. This research validates our usage of RF techniques in the project.

2.6 Assessing Models with ROC Curves and Confusion Matrix

Rekha (2008) employed several DM techniques for fraud analysis and illustrated how to interpret a confusion matrix and how ROC curves can be used for more instinctive model evaluation. This research paper is a valuable resource, allowing our team to comprehensively evaluate models from diverse perspectives.

3. Data Understanding and Preparation

3.1 Data Understanding

World Plus provides a dataset with 220,000 records of historical customer data from a previous product offering. The dataset includes 15 explanatory variables, four numeric and 11 categorical variables, and a target variable indicating whether the customer purchases the product (1) or not (0).

3.2 Data Preparation

For data preparation, the "ID" variable is excluded due to its lack of impact on the target variable. Of 220,000 records, 8.3% (18,268 instances) in the "Credit_Product" variable have missing values, reflecting customers' responses about active credit products. Due to the potential sensitivity of this question, we consider it MNAR, implying that not responding may form a distinct category. Following Zhuchkova and Rotmistrov (2022), CCA or MIM are suitable under MNAR. However, CCA alters the "Target" variable distribution (14.8% to 9.2%, as presented in Figure 1). By choosing MIM, we retain the entire dataset, interpreting missing values as "not_respond," preserving the original target variable proportion for unbiased analysis.

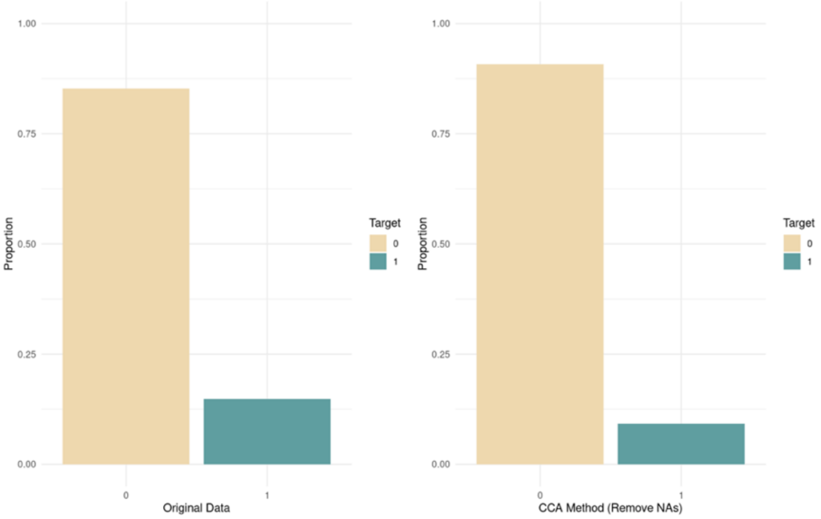

Figure 1. A comparison of the “Target” variable distribution

The "Dependent" variable anomalies (-1 responses) are removed. No outliers are observed in continuous variables. Categorical variables undergo encoding: "Account_Type" (label-encoded), "Active" (binary), and "Occupation" (one-hot). We employ stratified sampling to sample 40% of the dataset, maintaining proportional target values (0 and 1). This subset, representative of the original data, is split into 80% training (for model building) and 20% test (for model evaluation) sets.

The target distribution is imbalanced (85.2% for target=0, 14.8% for target=1, a 6:1 ratio as in Figure 1). Following Wongvorachan et al. (2023)'s recommendation for extremely imbalanced classification, our team apply a hybrid resampling approach using the "both" method in the ROSE (Random Over Sampling Examples) package. We specify p=0.5 to equalise the minority and majority groups to enhance the performance of predictive models.

4. Modelling

As the company's goal of managing marketing costs, the model could identify the group most responsive to the marketing campaign. Moro et al. (2014) and Apampa (2016) stated that LR, DT, and RF are great classification models that perform well in predicting categories. However, NN and SVM have more flexibility since they can learn simple and complex patterns. Consequently, we use these five models for analysing World Plus data, not only due to the similarity between the business field analysed in the research paper and World Plus but also because of these models' variability, which allows us to test the data from multiple perspectives.

We first use feature selection on our training data:

- As suggested by Yu and Abdel-Aty (2013), the CART method is used for SVM model building, showing that Registration and Credit_Product features are more decisive than others (Figure 2).
- Information Gain (IG) is used for DT and RF models, which shows that Years_at_Residence, Occupation_Other, and Account_Type have zero IG (Figure 3); these columns are removed for building the two models.

Furthermore, the NN model has a limitation in that it can use only numerical variables, and these data need to be normalised, so only three variables are used, which are "Age", "Years_at_Residence", and "Vintage"; and we use all variables (except "ID" variable) to build LR model.

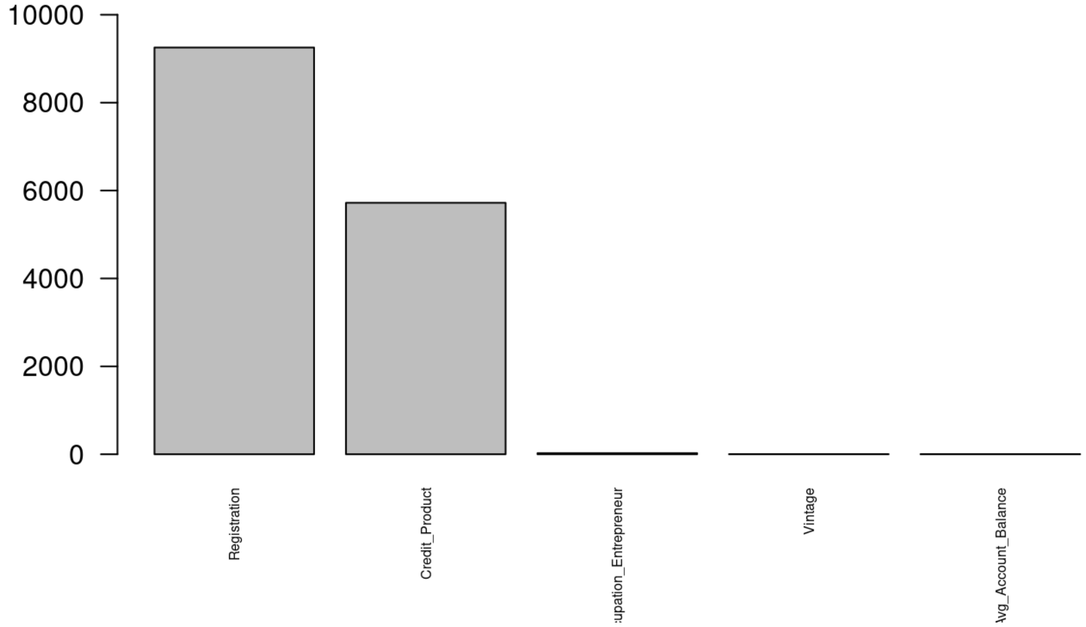

Figure 2. Results of most explanatory variables from CART

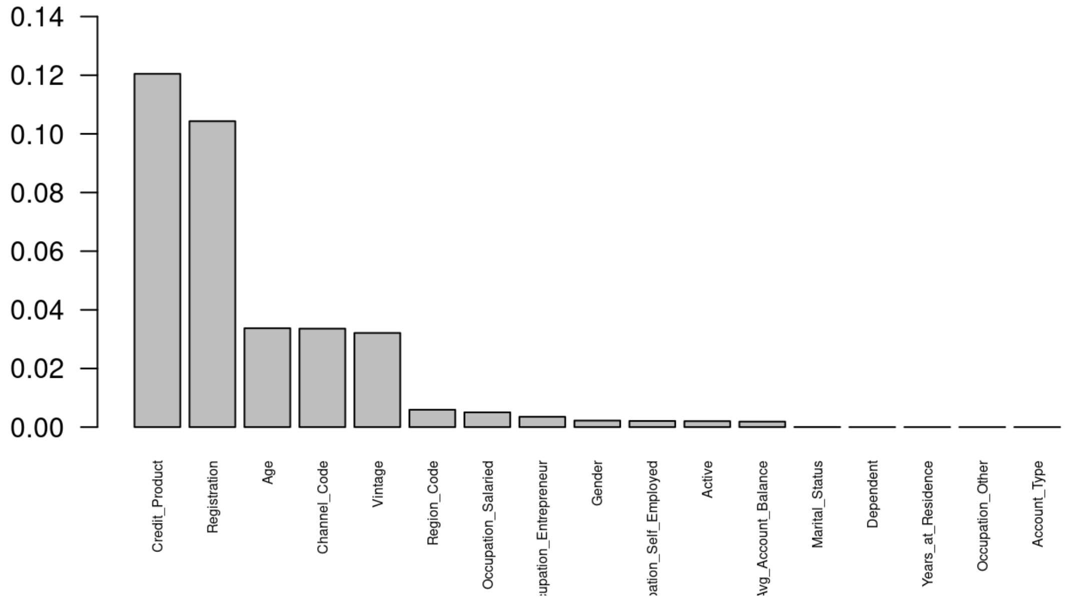

Figure 3. Results of most explanatory variables from Information Gain

The five DM mentioned above algorithms are used to predict the target variable. RStudio was used to build the models, and their related functions and features are summarised in Table 1.

Selected Models	Function in R Studio
SVM	svm() function is used with the kernel as “radial”, as data is not linearly separated, scale as “TRUE”, and probability as “TRUE” to obtain probability class.
DT	ctree() is used with its default setting.
RF	randomForest() is used with its default setting.
LR	glm() function is used with family as "binomial" due to its binomial distribution.
NN	neuralnet() function is used with the numerical variable as input to predict outcome probabilities. Rounding half of the numerical inputs determines the number of hidden nodes.

Table 1. Modelling

After obtaining the results, we further tuned SVM, DT, and RF models by adjusting their parameters. However, there are no tuneable hyperparameters in glm() function, and another LR function, glmnet(), requires all-numeric columns to tune, which does not applicable to the dataset.

5. Evaluation and Deployment

5.1 Model Evaluation

The evaluation of the models is based on figures within the confusion matrix (accuracy, precision, recall, and F1 score), ROC curves, and comparing the AUC value (Area Under the ROC Curve), as Rekha (2008) suggested. Aiming for cost minimisation, we select the model based on the precision rate, as precision is a perfect measurement of whether False Positive's cost is too high. After analysing five different modelling algorithms, the most accurate and best model for our problem is the RF model after tuning (Figure 5).

Metrics	Logistic Regression	Decision Tree	Random Forest	SVM	Neural Network
Accuracy	0.8737	0.8335	0.8830	0.8621	0.5452
Precision	0.5491	0.4648	0.5743	0.5210	0.2224
Recall	0.8096	0.8414	0.8032	0.8196	0.8329
F1	0.6544	0.5988	0.6697	0.6371	0.3511
AUC	0.9167	0.9054	0.9190	0.8818	0.6783

Table 2. Model Performance before tuning

Before the tuning process, the RF model shows that it is superior to others. Its accuracy and precision are the highest values, just as He et al. (2018) proposed. Furthermore, its F1 score of 0.6697 reflects a well-balanced precision and recall, making it a good model overall. Also, its AUC value indicates a better capability to discriminate between positive and negative instances.

Metrics	Decision Tree	Random Forest	SVM
Accuracy	0.8452	0.8959	0.8614
Precision	0.4861	0.6170	0.5196
Recall	0.8396	0.7780	0.8214
F1	0.6157	0.6882	0.6365
AUC	0.9112	0.9194	0.8789

Table 3. Model performance after tuning

In model tuning, tuned RF performs better than untuned one (Figure 4) and the best among others (Figure 5) in all metrics except the recall rate, as shown in Table 3.

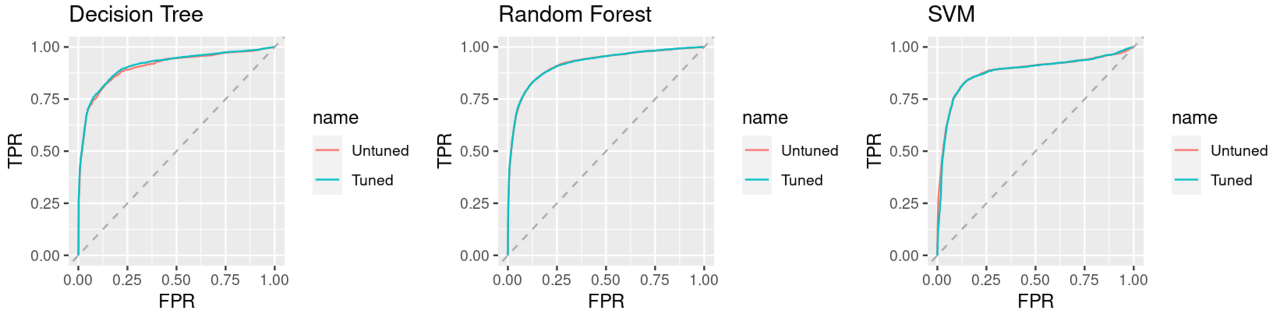

Figure 4. ROC curve comparison of tuned and untuned models

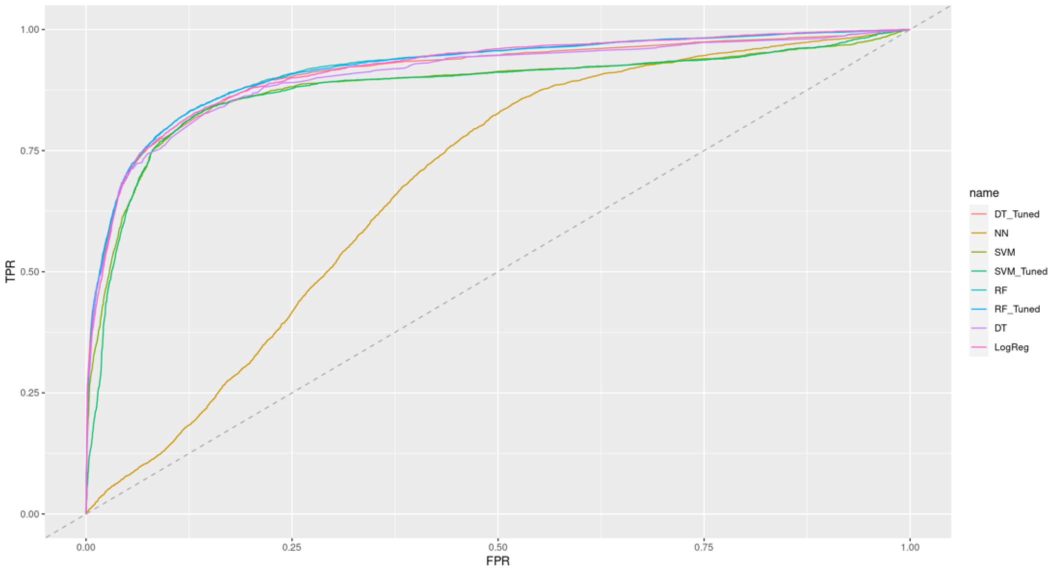

Figure 5. ROC curve comparison for all models

Expected Value

We use an expected value measure to compute the expected profit of eight models. We thus propose employing the RF model after tuning, which helps mitigate the marketing cost associated with false positives. On average, a financial institution can generate £19.12 profit on top of every pound spent on marketing. For the current situation, we assume that World Plus reaches out to all customers without the predictive model. The expected profit for every pound spent would be £3.68. With our RF model after tuning, World Plus can significantly boost profit per pound paid, £14.96, by reaching out exclusively to those predicted as Positives, thereby eliminating unnecessary costs (Appendix 7.2).

5.2 Deployment

A prediction system can be deployed by integrating the RF model results into the existing customer relationship management (CRM) or marketing automation platforms. Furthermore, to maintain the efficiency of the machine learning model, regular monitoring and maintenance after implementation are important (Studer et al., 2021); updates based on evolving data accumulated in the future will further ensure sustained accuracy of the model in the long run.

6. Conclusion, Recommendation, and Limitation

6.1 Conclusion

To deal with the company’s demands for precise prediction and better resource management, employing lead conversion models can help the bank avoid unnecessary expenses. This report explores models built through five algorithms evaluated based on three measures. Finally, our team selects the tuned RF model as the best to employ due to its high accuracy, precision, F1 score, and overall AUC.

6.2 Recommendation

Moreover, we observed that customers who visited the bank for product registration were more likely to convert into leads. By promoting initiatives that encourage customers to visit the bank for product registration, the probability of product conversion after receiving a marketing campaign could be further improved.

However, even though NN is considered the best predictive model in the banking industry, according to Moro, Cortez, and Rita (2014 and 2015), the model needs multiple variables to be constructed. Hence, an even better predictive model can be made if we increase the amount of numerical and normalised data for NN model construction.

6.3 Limitation

Finally, there are a few limitations to our procedures. The model requires a particular amount of time to generate a predictive result and high-quality hardware to run this model.

7. Appendix

7.1 Data Dictionary

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

7.2 A comparison of the expected profit per pound spent after using each model:

Assumptions: Profit over every £1 spent on Marketing campaign per customer is £19.12[1]

Criteria/Model	Actual data[2]	RF	RF (tuned)	DT	DT (tuned)	SVM	SVM (tuned)	LR	NN
Probability of right targeted customer prediction	14.80%	76.44%	78.10%	70.92%	72.12%	74.10%	74.01%	75.41%	64.61%
Probability of wrong targeted customer prediction	85.20%	2.91%	3.28%	2.34%	2.37%	2.66%	2.64%	2.81%	6.47%
Expected Value (£)	3.68	14.64	14.96	13.58	13.81	14.19	14.17	14.44	12.41

From the table, the RF model, after tuning, has the highest rate of predicting World Plus targeted customers at an accuracy of 78.10%. In monetary terms, RF model, after tuning, can generate an average profit of £14.96 on top of every pound spent on marketing campaigns.

[1] According to Cocheo S. (2019) and Daher M. and Kneer C. (2022), we imply that World Plus' total asset is less than US 500 million

[2] We assume that the company applies the Marketing campaign to all customers.

8. Reference

Apampa, O. 2016. “Evaluation of classification and ensemble algorithms for bank customer marketing response prediction”, *Journal of International Technology and Information Management*, 25(4), 6.

Bhowmik, Rekha. 2008, "Data Mining Techniques in Fraud Detection," Journal of Digital Forensics, *Security and Law*: Vol. 3: No. 2, Article 3.

Cocheo S. (2019), “How big should bank marketing budgets be for profitability & growth?”. https://thefinancialbrand.com/news/bank-marketing/bank-marketing-budgets-advertising-roi-strategy-88835/ (Accessed: 1 December 2023).

Daher M. and Kneer C. (2022), “Mountains of debt and investment flows: what can we learn from SMEs’ investment behaviour during and after the global financial crisis?”. https://www.bankofengland.co.uk/-/media/boe/files/financial-stability-paper/2022/mountains-of-debt-and-investment-flows.pdf (Accessed: 1 December 2023).

Lingjun, He; Levine, Richard A.; Fan, Juanjuan; Beemer, Joshua; and Stronach, Jeanne. 2019, “Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research," *Practical Assessment, Research, and Evaluation*: Vol. 23, Article 1.

Moro, S., Cortez, P. and Rita, P., 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62,* pp.22-31.

Moro, S., Cortez, P. and Rita, P., 2015. “Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing Campaigns”. *Neural Computing and Applications*, *26*, pp.131-139.

Rongjie Yu, Mohamed Abdel-Aty, 2013, “Utilizing support vector machine in real-time crash risk evaluation”, *Accident Analysis and Prevention, 51, pp. 252– 259*

Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S. and Müller, K.R. (2021) 'Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology', *Machine Learning and Knowledge Extraction*, 3(2), pp.392-413.

Wongvorachan, T., He, S. & Bulut, O. 2023, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining", *Information*, vol. 14, no. 1, pp. 54.

Zhuchkova, S. & Rotmistrov, A. 2022, 'How to choose an approach to handling missing categorical data: (un)expected findings from a simulated statistical experiment', *Qual Quant*, vol. 56, pp. 1–22.