WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programme

Student Number:	**2053845, 289524, 5587205, 5562214, 5558022, 5528636**
Module Code:	**IB9BW0**
Module Title:	**Analytics In Practice**
Submission Deadline:	**06/12/2023**
Date Submitted:	**06/12/2023**
Question:	**Question: A mid-size private bank, "World Plus", requested you to deliver a pitch to win a major contract with them to develop and deploy a lead prediction system.**
Word Count:	**1971 (exc. Figures, Tables, Headings, Captions)**
Number of Pages:	**21**
Have you used Artificial Intelligence (AI) in any part of this assignment?	**No**

Academic Integrity Declaration

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that the work is all my own, except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I

have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.

- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

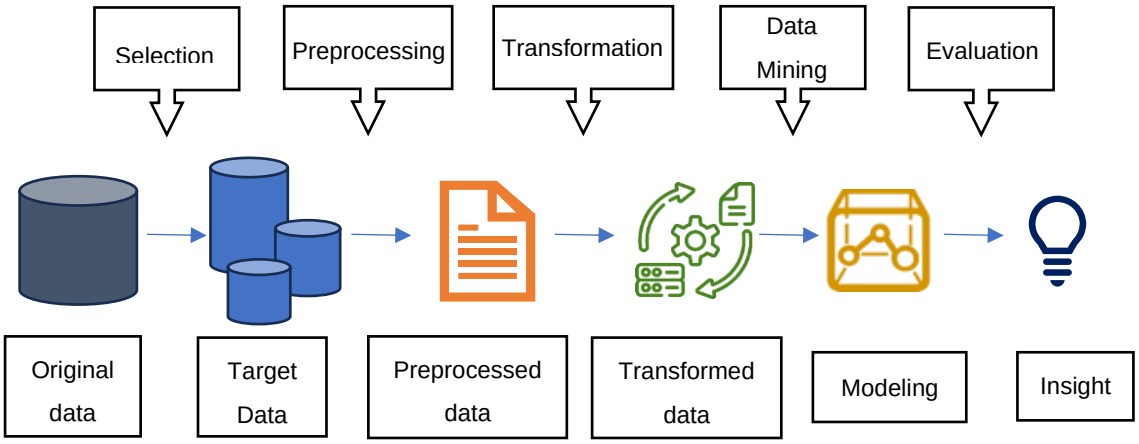
Table of Contents

Business Understanding

World Plus, a mid-size private bank, confronts the challenge of precisely identifying potential leads for their new term deposit product, resulting in inefficient marketing expenditure. The proposed solution entails the development and deployment of a Lead Prediction System using advanced data mining techniques on a dataset of 220,000 customer records. The objective is to curtail unnecessary expenses by strategically targeting prospective customers who will convert and buy the new term deposit.

Problem-solution Approach: CRISP-DM

We will be using an industry wide approach “Cross-Industry Standard Process for Data Mining (CRISP-DM)” for the project. Figure below demonstrates a step-by-step procedure of the analysis.



Process flow diagram

Data Understanding and Data Preparation

The dataset provided by the client contains customer features and a target variable, which relate to the purchase (1/0) of a similar term deposit product.

The variables in the dataset and its description are as presented in table 1:

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	Duration in the current residence (in years)
Occupation	Occupation type of the customer
Channel_Code	Acquisition channel code used to reach the customer
Vintage	Number of months that the customer has been associated
Credit_Product	If the customer has any active credit product
Avg_Account_Balance	Average account balance for the customer in the last 12 months
Account_Type	Account type of the customer: Silver, Gold, Platinum
Active	If the customer is active in the last 3 months
Registration	Whether customer has visited the bank for offered product registration
Target	Whether the customer has purchased the product (1 = purchase, 0 = no purchase)

Table 1

Before the initial model creation, dataset cleaning was imperative. Commencing with removing the 'ID' column, which solely functioned as a unique identifier without meaningful information for each row, redundant data was eliminated. Addressing the 18233 missing values in the "Credit_Product" column is pivotal. The primary strategy refrained from outright removing cases with missing values due to concerns about potential information and statistical power loss, particularly for the minority class.

Concurrently, investigating the impact of such an imputation method, Kwak and Kim (2017) conducted research on the impact of various imputation techniques on data analysis. The authors cautioned against this "Trimming" approach, emphasising its impact on reducing sample size and representing crucial data points. Therefore, a negative impact on model performance is expected. Nevertheless, for comparative analysis, this dataset was included.

Another technique the study mentions is Imputation Analysis, which involves replacing missing values with substituted values derived from statistical analysis, creating a more complete dataset for analysis. Given their mention of using explicit modelling approaches to estimate the parameters of each distribution, the decision was made to handle instances labelled as "no" in the Credit_Product column by assigning them as '0'. Simultaneously, "yes" instances were assigned with the value '1'. Subsequently, we computed the mode of the Credit_Product distribution and utilised this modal value (0) to replace the missing values within the column. Nevertheless, this approach relies on the presumption that the absent values adhere to the identical distribution as the observed ones. This assumption could skew subsequent analyses, resulting in inaccurate predictions.

Given the disadvantage of the two data-cleaning approaches, further research introduced MICE (Multivariate Imputation by Chained Equations). MICE's fundamental idea revolves around generating multiple plausible values for each missing data point, utilising other variables as predictors and creating multiple complete datasets (Meghanadh et al., 2018). This paper supplemented the cleaning process with a predictive solution for imputation prior to modelling. The preference for MICE over Single Imputation Methods is grounded in its capacity to address uncertainty and avoid variance underestimation, as elucidated by Rubin (1987). The

research supports our strategy of creating multiple datasets to enhance flexibility and to allow consideration of various results before reaching a conclusive decision. In all approaches, a precautionary measure involved duplicating the original dataset to maintain a distinct version suitable for imputation.

Following data partitioning to segregate the dataset into test and training sets, a significant imbalance was identified in the target variable. Instances where customers did not purchase the product (class 0) were notably more prevalent than class 1 in the target variable. To address the imbalanced target variable, a decision was made to undersample the majority class and oversample the minority class, as Batista et al. (n.d) suggested. This approach establishes better-defined class clusters, preventing the development of models that prioritise accuracy on the majority class while performing inadequately on the minority class.

The objective of balancing the class distribution is twofold: it enhances the model's capacity to identify patterns and make accurate predictions for both classes and fosters generalisation to unseen data, thereby increasing robustness in real-world scenarios. Nevertheless, a significant drawback associated with under-sampling is acknowledged – the potential loss of valuable data crucial for learning. Consequently, over-sampling methods have been favoured, as they tend to yield more accurate results by directly addressing the lack of minority class examples.

Data Visualisation

Figure 1 allows us to compare the account balances distribution for customers who purchased and did not purchase the target product. It can be observed that the distribution is skewed to the right rather than being normally distributed. This means the majority of customers have relatively lower average account balances, contributing to the higher frequency on the left side of the distribution, but a very small proportion outliers having significantly higher than average account balance. Due to this, we used the median to gauge the central tendency of account balance, which was found to be 885,069.

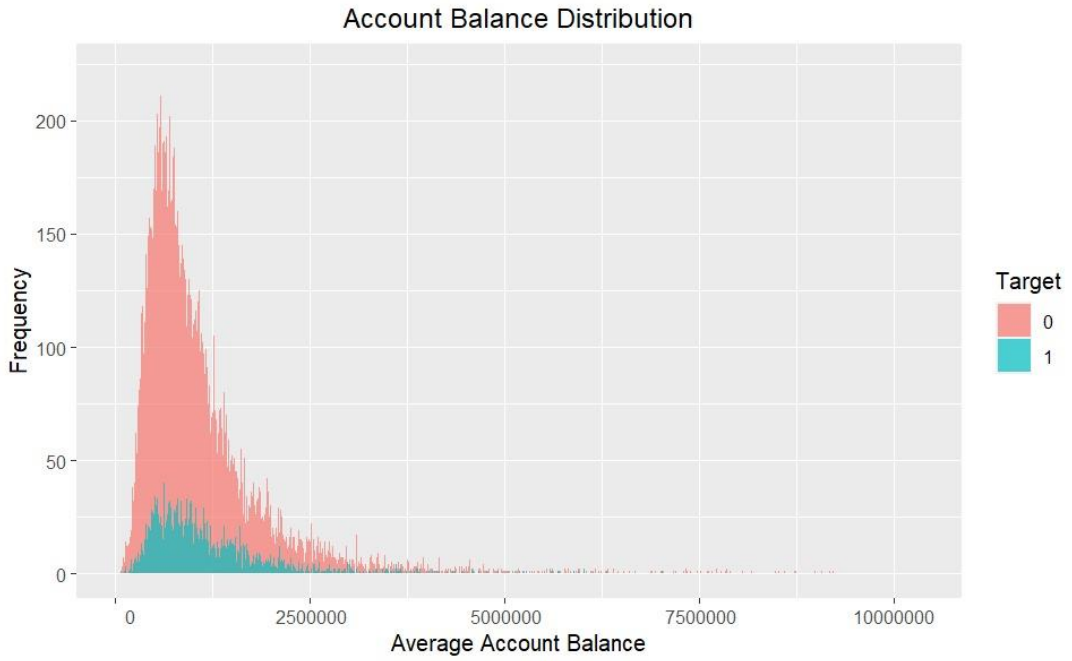


Figure 1

Figure 2 implies customers with credit product have a higher conversion rate (28%) compared to those who don't have credit product (9%), suggesting that targeting marketing efforts towards the former could be more viable.

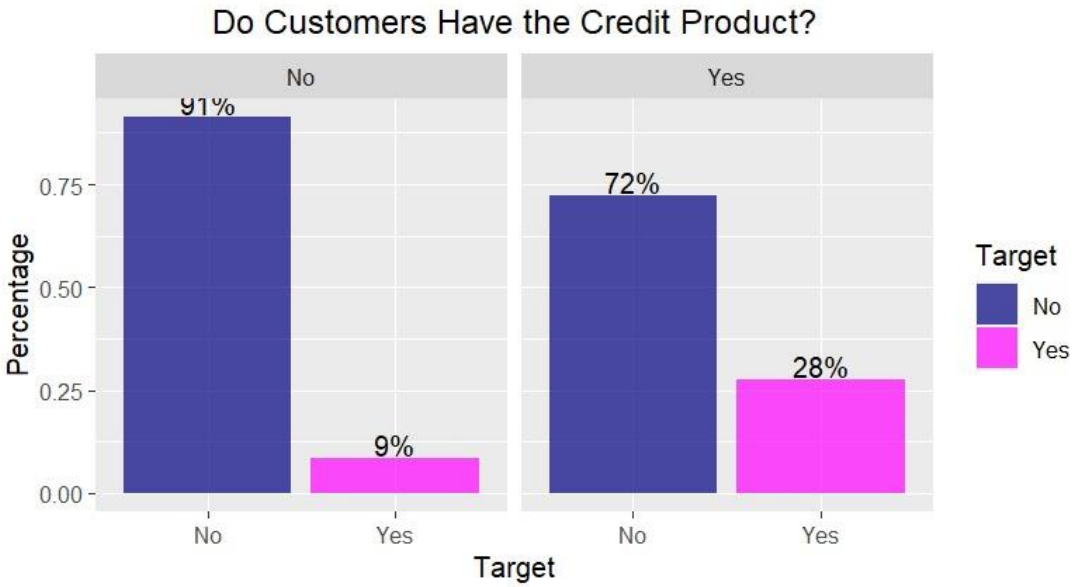


Figure 2

Examining Figure 3, it is evident that a significant portion of positive outcomes in the 'Target' variable originates from customers categorised as Entrepreneurs. This observation leads us to anticipate that Entrepreneurs are more prone to conversion.

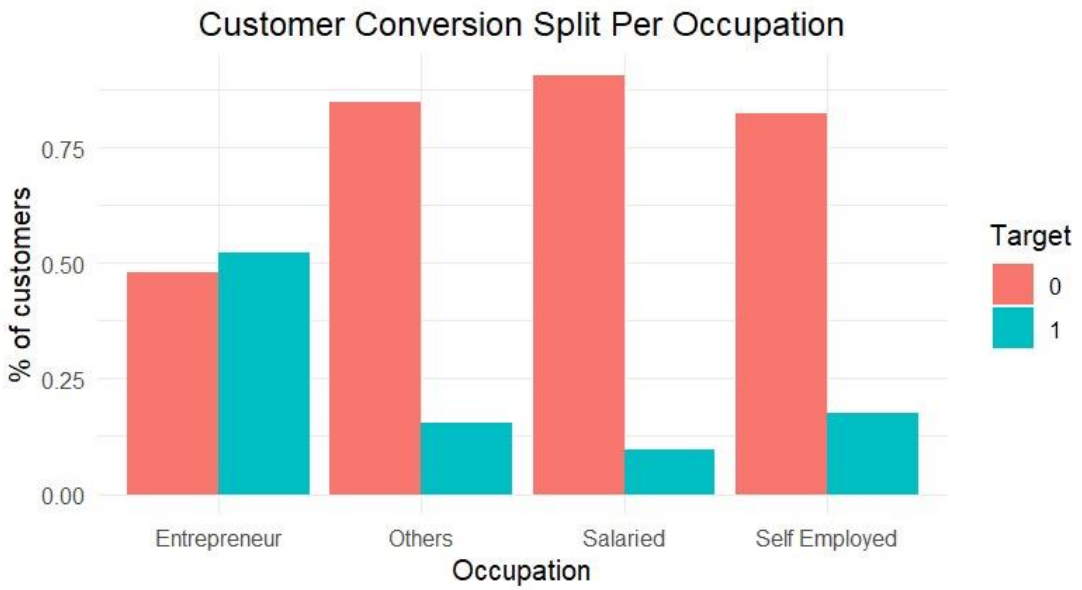


Figure 3

Modelling Techniques

From Schwenke and Schering (2014), we are introduced to standardised language for discussing the performance metrics of the predictive model. True positives (TP) denote instances in which the model accurately predicts customer conversion. Conversely, false positives (FP) denote inaccuracies, wherein customers are incorrectly identified as likely to convert when they do not. Conversely, false positives (FP) denote inaccuracies, wherein customers are incorrectly identified as likely to convert when they do not. Analogously, the article also states that the terms True negatives (TN) and false negatives (FN) are applicable in a similar context.

Drawing upon these metrics, the creation of confusion matrices allows a comparative analysis of various evaluation measures, encompassing precision (TP/TP+FP), recall (TP/TP+FN), accuracy, and F1-score.

Precision measures the accuracy of predicting customer purchases, reducing operational costs by ensuring precise predictions. Recall evaluates the model's ability to identify all actual purchasing customers, aiming to capture as many true positives as possible and minimizing the risk of overlooking opportunities.

In addressing imbalanced data, Slamet et al. (2023) underscore the reliability of the F1-score in comparison to precision and recall. Consequently, the primary focus centred on the F1-score, representing the harmonic mean of both precision and recall. The evaluation proceeds with the construction of Receiver Operator Characteristic (ROC) curves for the models, followed by a direct comparison of the areas under these curves (AUC). AUC, a metric ranging from 0 to 1, was calculated to demonstrate its clear indication of model performance, as emphasised by Hamel (2009), where a higher AUC value corresponds to superior performance.

Hamel's research underscores the limitations of confusion matrices, particularly their sensitivity to data anomalies, especially class skew. Utilising ROC curves addresses this issue and provides visual analysis.

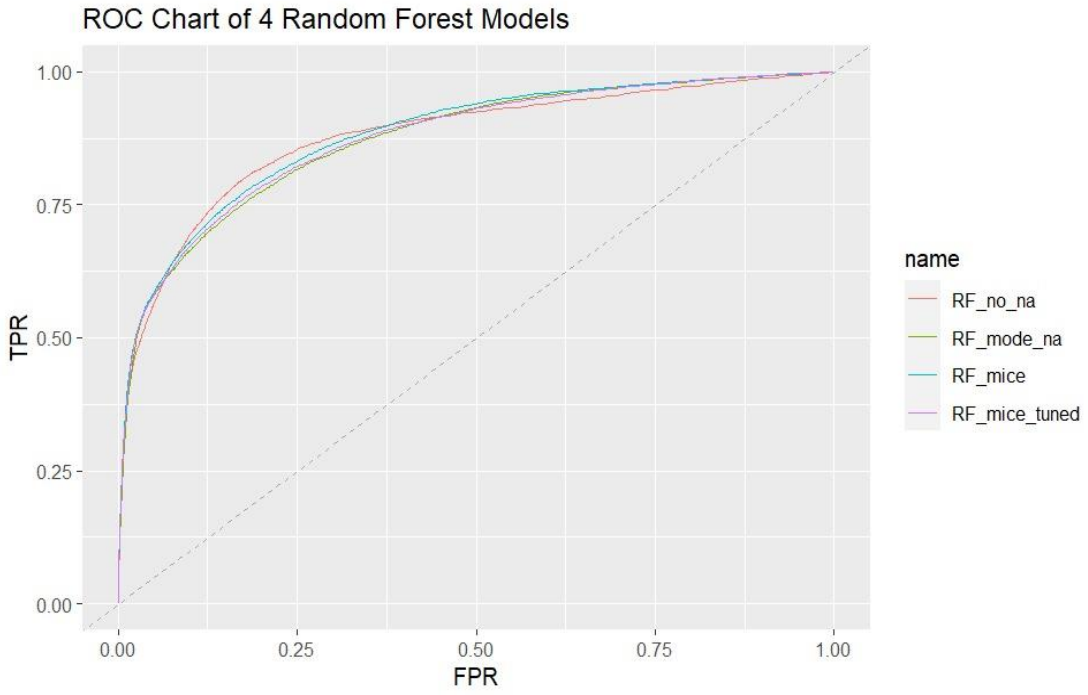


Figure 4

The models for each dataset were built and tested for class prediction. The initial focus was on the Random Forest (RF) model. Registration emerged as a significant factor influencing the target variable, indicating that registered customers are more likely to be prospective. The confusion matrix highlighted that the model trained on the MICE dataset achieved the highest F1-score at 0.5933. Post-tuning, the F1-score and precision improved to 0.6032 and 0.58031, respectively, albeit decreasing recall.

However, creating the ROC chart (Figure 4), the tuned model had lower AUC of 0.8751 compared to 0.8814 of initial model.

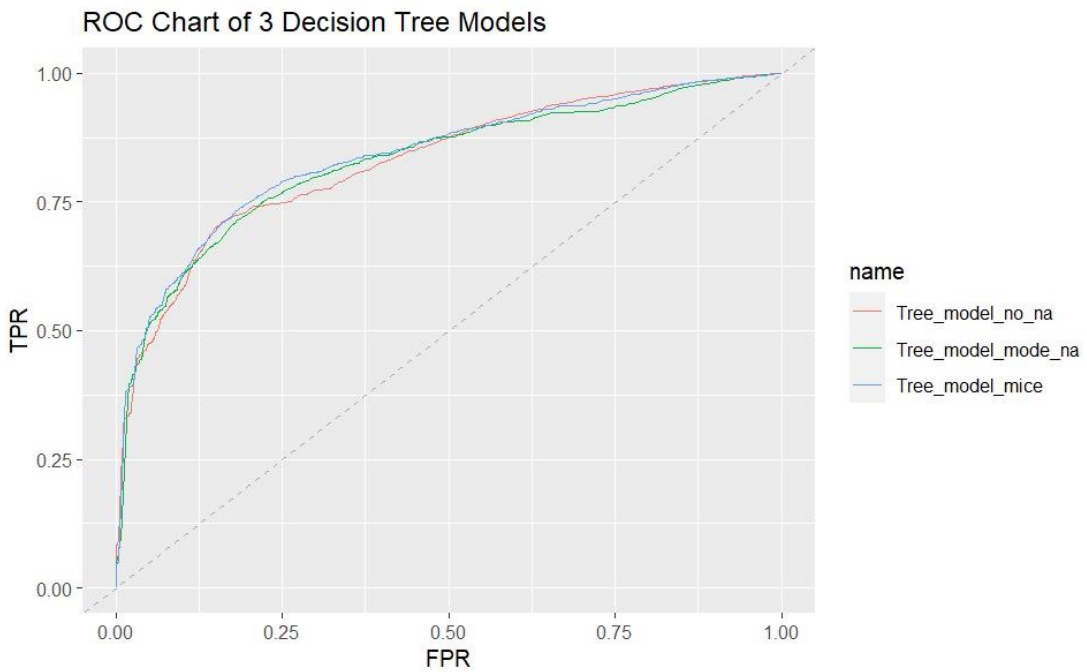


Figure 5

Transitioning to the second model, the Decision Tree model, a preliminary step involves computing information gain and deleting attributes with non-positive IG to mitigate the impact of irrelevant variables. According to Delen, Kuzey and Uyar (2013), C5.0 provides the highest accuracy among popular decision tree algorithms. This paper supplemented the choice to build the Decision Tree model using the c5.0() function.

The confusion matrix is computed revealing that the model using the MICE dataset is chosen as it achieves the highest F1-score of 0.5164. The ROC chart (Figure 5) further highlights this superiority, displaying the largest AUC at 0.8374.

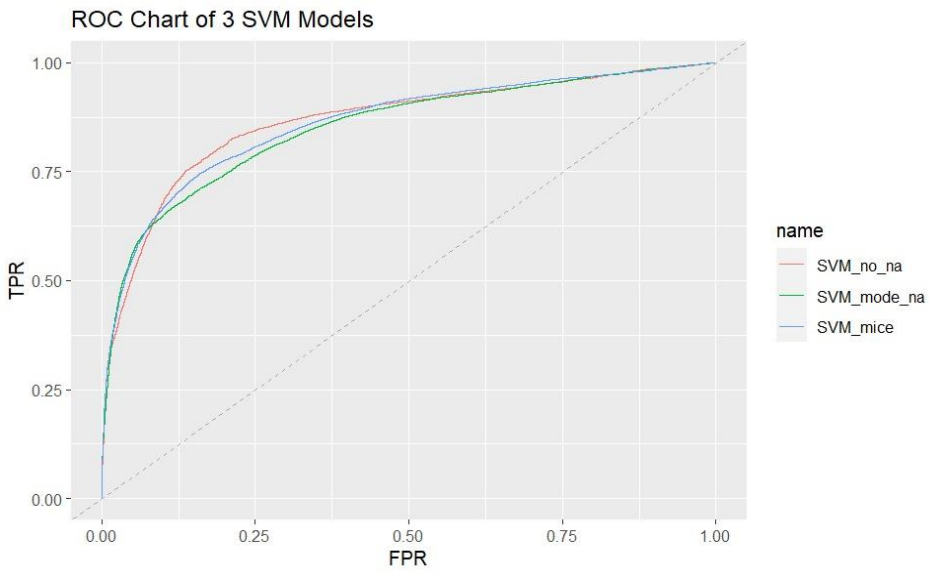


Figure 6

Another approach included the use of the Support Vector Machine (SVM) Model, as according to Shin et al. (2005), SVM captures geometric characteristics of the feature space without deriving the weights of networks from the training data.

It is also mentioned that back-propagation neural network also performs well in pattern recognition tasks, although this method has difficulty in finding an acceptable model structure and optimal solution. The study was conducted to predict bankruptcy and was utilised for its financial implications within our business context.

According to the confusion matrix, the model using mice dataset achieves the highest F1-score of 0.5653. However, the model that removes missing values exhibits the greatest AUC of 0.8659, which is slightly greater than 0.8616 of the MICE dataset. To strike a balance between F-score and AUC, the choice is made to opt for the model using the MICE dataset.

Furthermore, Shin et. al also found that the smaller training data size, the better generalisation performance of SVM. Hence a 30% subset was taken from the initial

training data. To address non-linearity in training datasets, the kernel was set to radial.

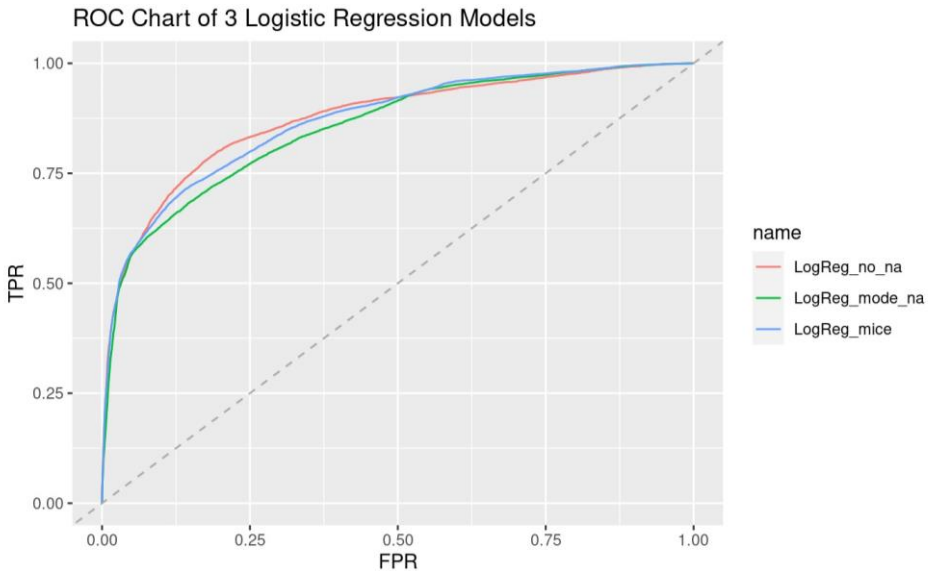


Figure 7

The fourth model under consideration is the Logistic Regression model. The variant that removes missing values shows the highest AUC of 0.873, slightly edging out the one utilising the MICE dataset, which has an AUC of 0.8693 (Figure 7). However, the former model has a significantly lower F1-score whereas the MICE dataset-based model achieves the highest F1-score of 0.5518. Therefore, the choice leans towards the model using the MICE dataset.

Model Evaluation

	bestRF	RF_model_mice	treemodel_mice	SVM_mice	LogReg_mice
F1-score	**0.60791**	0.5932	0.5164	0.5653	0.5518
AUC	**0.8751**	0.8814	0.8374	0.8616	0.8693
Precision	**0.58031**	0.5134	0.394	0.4595	0.4425

Table 2

In the comparison provided, the selection encompasses the five top-performing models. The evident preference for models based on the MICE dataset underscores the success of integrating MICE as a strategic approach in the research. The superiority of the two RF models, both in terms of F1-score and AUC, stands out prominently. Schonlau and Zou (2020) assert that, particularly when confronted with sizable datasets, RF outperform alternative modelling approaches, as illustrated in their analysis of credit card default data—a scenario particularly relevant to the business context under consideration.

RF's capability to handle both continuous and categorical predictors is emphasised, along with its effectiveness in mitigating overfitting by leveraging diverse tree-based models and averaging their predictions. However, this may come at the cost of interpretability, given the generation and aggregation of multiple tree models. Additionally, the system is designed with objective of aiding the bank in reducing unnecessary expense by targeting prospective customers. The goal is to ensure that every individual identified as a prospective customer through predictions eventually becomes a future purchaser of the new product.

Thus, precision should be also considered. RF models show higher precision over other types of models. Ultimately, the tuned RF model is better than the original RF

model in this sense because of the substantial precision increase from 0.5134 to 0.58031, not to mention the increased F1-score.

An alternative approach for comparing models involves considering expected values, derived by using the expected profit for each outcome from the confusion matrix, and multiplying it by its respective probability. Secondary research is conducted to obtain the profit derived from selling term deposits and an average marketing cost per customer. According to Fitch Ratings (2023), the industry average profit margin is 2.2%. Additionally, Baylin (2023) estimates the marketing cost per customer within the financial industry as approximately $500. Utilising these figures, the expected value for each outcome in the confusion matrix can be calculated accordingly.

1) **RF mice tuned model**

Confusion Matrix: Mice dataset

Model	**RF Tuned**		**Decision Tree**		**Logistic Regression**		**SVM**	
Prediction	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
0	51725	3524	45003	2447	47230	2604	48258	2668
1	4497	6218	11219	7295	8992	7138	7964	7074

Table 3

RF mice tuned model	**Amount**	**Probability**
Profit	1958	
Cost	500	
True positive	1458	0.09
False negative	0	0.05
True negative	0	0.78
False positive	-500	0.07
Expected profit per customer	**103.35**	

Table 4

2) **Decision Tree mice model**

Decision Tree mice model	Amount	Probability
Profit	1958	
Cost	500	
True positive	1458	0.11
False negative	0	0.04
True negative	0	0.68
False positive	-500	0.17
Expected profit per customer	**76.2**	

Table 5

3) **Logistic Regression mice model**

Logistic Regression mice model	Amount	Probability
Profit	1958	
Cost	500	
True positive	1458	0.11
False negative	0	0.04
True negative	0	0.72
False positive	-500	0.14
Expected profit per customer	**89.61**	

Table 6

4) **SVM mice model**

SVM mice model	Amount	Probability
Profit	1958	
Cost	500	
True positive	1458	0.11
False negative	0	0.04
True negative	0	0.73
False positive	-500	0.12
Expected profit per customer	**95.99**	

Table 7

5) **No model (Random selection)**

No model	Profit	Probability	Expected Value
Positive (1)	1458	15%	218.7
Negative (0)	-500	85%	-425
Expected profit per customer	**-206.3**		

Table 8

Conclusion

In conclusion, the empirical evidence underscores the ‘bestRF’ model as the optimal choice for integration into the lead prediction system. Furthermore, this model demonstrates the highest expected profit per customer at 103.35. Moreover, selecting a predictive model such as ‘bestRF’, the client now can target their marketing efforts with only those customers that show a high conversion propensity.

The analysis reveals that deviating to a random approach to customer selection, results in a financial loss from the perspective of expected value, which underscores the financial benefit of implementing the suggested model.

References

1. Bailyn, E. (2023) *Average Cost Per Lead by Industry 2024*. Available at: https://firstpagesage.com/reports/average-cost-per-lead-by-industry/ (Accessed: 4 December 2023)

2. Batista, G., Prati, R., and Monard, M. (n.d.) *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*. Available at: https://www.kdd.org/exploration_files/batista.pdf. (Accessed: 4 December 2023)

3. Delen, D., Kuzey, C., and Uyar, A. (2013) 'Measuring firm performance using financial ratios: A decision tree approach', *Expert Systems with Applications*, 40(10), pp. 3970-3983. Available at: https://doi.org/10.1016/j.eswa.2013.01.012 (Accessed: 4 December 2023)

4. Fitch Ratings (2023) *Many DM100 Bank Net Interest Margins Have Peaked in 2023*. Available at: https://www.fitchratings.com/research/banks/many-dm100-bank-net-interest-margins-have-peaked-in-2023-27-11-2023 (Accessed: 4 December 2023)

5. Hamel, L. (2008) 'Model Assessment with ROC Curves', *ResearchGate*. Available at: https://www.researchgate.net/publication/314417632_Model_Assessment_with_ROC_Curves (Accessed: 5 December 2023)

6. Kwak, S. G., and Kim, J. H. (2017) 'Statistical data preparation: management of missing values and outliers', *Korean Journal of Anesthesiology*, 70(4), pp. 407–407. Available at:

https://doi.org/10.4097/kjae.2017.70.4.407 (Accessed: 4 December 2023)

7. Meghanadh, B., Aravalath, L., Joshi, B., Sathiamoorthy, R., and Kumar, M. (2019) 'Imputation of Missing Values in the Fundamental Data: Using MICE Framework', *Journal of Quantitative Economics*, 17(3), pp. 459-475. Available at: https://EconPapers.repec.org/RePEc:spr:jqecon:v:17:y:2019:i:3:d:10.1007_s40953-018-0142-7 (Accessed: 4 December 2023)

8. Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley

9. Schonlau, M. and Zou, R.Y. (2020) 'The random forest algorithm for statistical learning', *Sage Journals*, 20(1). Available at: https://doi.org/10.1177/1536867X20909688 (Accessed: 4 December 2023)

10. Schwenke, C., and Schering, A. (2014) 'True Positives, True Negatives, False Positives, False Negatives', *Wiley StatsRef: Statistics Reference Online*. Available at: https://doi.org/10.1002/9781118445112.stat06783 (Accessed: 4 December 2023)

11. Shin, K., Lee, T.S. and Kim, H. (2005) 'An application of support vector machines in bankruptcy prediction model', Expert Systems with Applications, 28(1), pp. 127-135. Available at: [https://doi.org/10.1016/j.eswa.2004.08.009] (https://doi.org/10.1016/j.eswa.2004.08.009) (Accessed: 4 December 2023)

12. Slamet, R. et al. (2023) 'Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification', *International Journal of Advanced Computer Science and Applications*, 14(6). doi:10.14569/IJACSA.2023.01406116.