Table of Contents

1 Introduction

In the world of big data, organisations strive to gain a competitive edge by developing decision support systems that help make better decisions. As companies navigate through information generated daily, predicting customer behaviour stands out as a big challenge. Predictive modelling has revolutionised business approaches, enabling companies to anticipate customer behaviour and optimise operations effectively. Its integration into decision support systems has become indispensable for businesses to compete within their industry.

Recognising this importance, World Plus, a mid-size private bank, is keen on implementing advanced lead predictive system. The objective is to accurately identify prospective lead customers while optimising marketing expenditures — a vital step for the bank's success, as optimising marketing and sales campaigns is a key issue under growing pressure to increase profits and reduce costs (Moro et al., 2014). This report aims to identify the most appropriate data modelling for World Plus using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework.

2 Literature Review

The CRISP-DM methodology puts forward an inclusive structure for executing data mining initiatives across sectors. Wirth and Hipp (2000) assessed the methodology using a response modelling trial to improve the effectiveness of email marketing. Their evaluation highlighted successful aspects of the approach, as well as some challenges encountered, which offered valuable insights on how to carry out our project as a group at each section of the framework.

To handle missing values in datasets, Anand and Mamidi (2020) evaluated the performance of Multiple Imputations using Chained Equations (MICE), an advanced form of Multiple Imputation that replaces missing values with what they could have been. Using real-world datasets, MICE consistently achieved superior performance for missing values under 10%, which influenced our decision to employ this approach for our dataset.

Blagus and Lusa (2013) explored SMOTE, a popular data sampling technique, on high-dimensional simulations and gene datasets. Contrasting metrics showed SMOTE failed to mitigate the bias except for some model classifiers such as kNN. SMOTE augmented variance and distance changes theoretically explained the poor performance. Because SMOTE is beneficial for low-dimensional tasks, random under-sampling outperformed SMOTE on key accuracy and AUC indices. Therefore, we have chosen to employ the under-sampling methodology.

Win and Bo (2020) developed a Random Forest model to effectively identify high-lifetime-value customers from transaction data. Feature selection was performed to choose the most relevant features for prediction. They leveraged this technique for its high performance in prediction tasks which produced high accuracy and recall in classifying the most profitable customers for retention efforts. This helped us understand the significance of hyperparameters in data modelling.

Saeed et al. (2022) highlighted the importance of an Exploratory Data Analysis (EDA). EDA recommends visually exploring the relationships between features and the target variable, as well as examining data imbalance. Additionally, they also suggest using classification metrics such as "Accuracy", "Precision", "Recall", and "F-measure" to compare the performance of different models. This report emphasises the importance of the above evaluation metrics in model evaluation, as it effectively maximises the profit by targeting as many clients as possible and minimises costs associated with non-targeted clients.

Kaur and Kaur (2020) demonstrated the methodologies of applying machine learning techniques in the banking sector for customer churn prediction, which is in line with our objective of building a lead predictive model for a private bank. Moreover, they reported the outstanding performance of Random Forest despite of imbalanced dataset and different sampling methods. These insights gave us extra attention to data balancing and the result of Random Forest while performing data preparation and evaluation respectively.

3 CRISP-DM Framework

CRISP-DM is a methodology that provides a structured approach to carry out a data mining project. It is an industry-independent process model which consists of six phases from business understanding to deployment.

3.1 Business Understanding

The business objectives for World Plus are defined as follows:

1. To accurately identify prospective customers who will purchase the new term deposit product.
2. To strategically target identified customers through appropriate communication channels to increase lead conversion.
3. To minimise unnecessary expenses and opportunity costs to improve sales and marketing operations.

The data modelling objectives for the team is defined as follows:

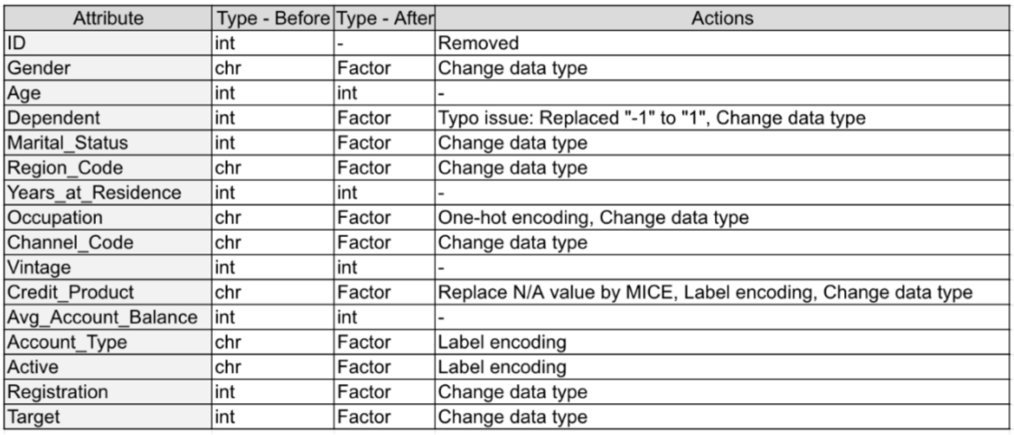
1. To produce a predictive modelling system that identifies customers who are more likely to purchase the new term deposit product.
2. To increase and keep a balance between precision and recall by minimising false positives and false negatives.

3.2 Data Understanding

The historic data collected during a similar product offering in the past consists of 220,000 instances with 16 attributes.

18,268 missing data within “Credit_Product” attribute was converted to either “Yes” or “No” using MICE. Since the missing data accounted up to 8.3% of the entire dataset, MICE was an appropriate choice to handle missing data (Mamidi and Anand, 2020).

Table 1 represents the attributes and the actions we took during the data pre-processing process.



Attribute	Type - Before	Type - After	Actions
ID	int	-	Removed
Gender	chr	Factor	Change data type
Age	int	int	-
Dependent	int	Factor	Typo issue: Replaced "-1" to "1", Change data type
Marital_Status	int	Factor	Change data type
Region_Code	chr	Factor	Change data type
Years_at_Residence	int	int	-
Occupation	chr	Factor	One-hot encoding, Change data type
Channel_Code	chr	Factor	Change data type
Vintage	int	int	-
Credit_Product	chr	Factor	Replace N/A value by MICE, Label encoding, Change data type
Avg_Account_Balance	int	int	-
Account_Type	chr	Factor	Label encoding
Active	chr	Factor	Label encoding
Registration	int	Factor	Change data type
Target	int	Factor	Change data type

Table 1: Attributes before and after data cleaning

3.2.1 Initial Analysis of the Dataset

After data cleaning, we explored the relationship of “Channel_Code” and "Registration” against customer’s purchase decision to better understand the current situation of World Plus.

Figure 1 indicates a misalignment between the communication channels and customer’s purchase decisions. X3 is the most effective channel for purchases, whereas X1 shows a significant performance gap. This illustrates there is a need to reconsider communication strategies and shift communication methods towards those demonstrating higher responsiveness such as X3.

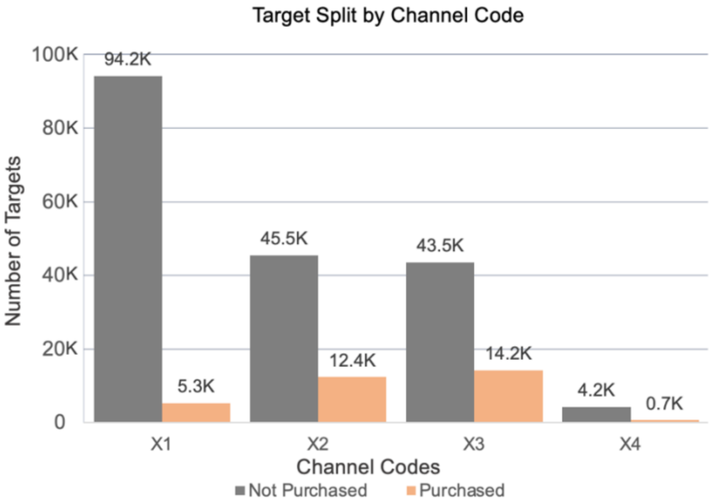


Figure 1: Relationship between target’s purchase decision and Channel Code

According to Figure 2, “Registration” has a significant impact on influencing the customers to purchase the product. The graph shows that more than half of the customers who came to the registration purchased the product. Therefore, a strong correlation is expected to be seen during data modelling.

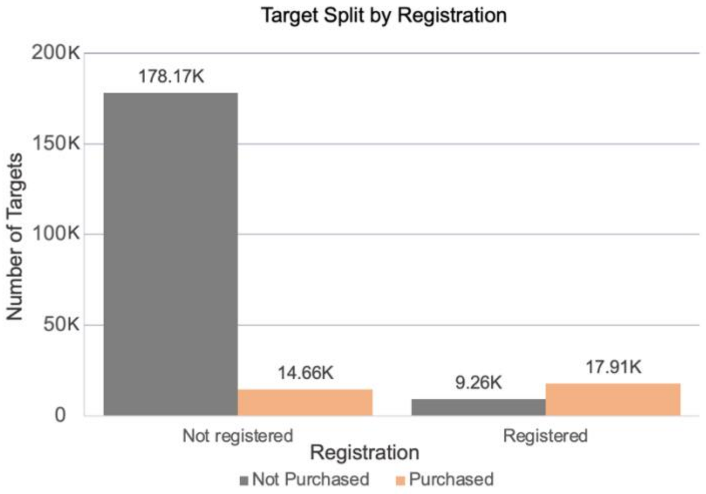


Figure 2: Relationship between target’s purchase decision and Registration

3.3 Data Preparation

The data preparation stage in the CRISP-DM framework involves structuring the data into a format ready for modelling to create the final dataset.

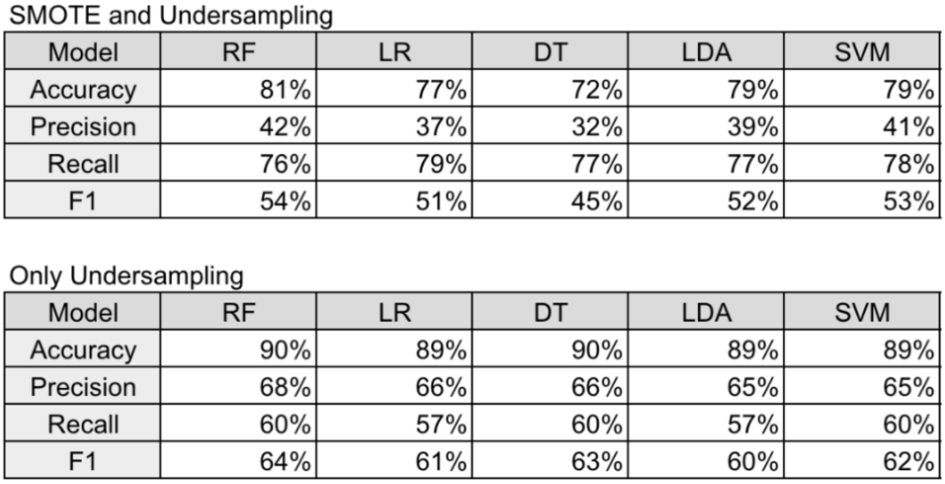
3.3.1 Data Partitioning and Data Balancing

70% of data is selected as training data and the rest 30% as test data.

In the next process, we tried two approaches to balance the dataset by under-sampling with SMOTE and then only using under-sampling.

The comparison of results with SMOTE and under-sampling shows the precision rate fell to approximately 30% for all models (Table 2). This is because applying SMOTE to achieve an equal balance with the majority class is not necessarily the best case for classifiers. Additionally, SMOTE is not beneficial for discriminant analysis classifiers in the low-dimensional setting, especially under-sampling (Blagus, R. and Lusa, L., 2013). Therefore, we changed the approach to use only the under-sampling technique with p=0.3, seed = 123 as shown in the Table 2.

In scenarios where certain features may predominantly characterise the majority class, under-sampling can lead to a more balanced representation of features, making the model more sensitive to features that are important for predicting the minority class (Lopez et al., 2013).



SMOTE and Undersampling

Model	RF	LR	DT	LDA	SVM
Accuracy	81%	77%	72%	79%	79%
Precision	42%	37%	32%	39%	41%
Recall	76%	79%	77%	77%	78%
F1	54%	51%	45%	52%	53%

Only Undersampling

Model	RF	LR	DT	LDA	SVM
Accuracy	90%	89%	90%	89%	89%
Precision	68%	66%	66%	65%	65%
Recall	60%	57%	60%	57%	60%
F1	64%	61%	63%	60%	62%

Table 2: Comparison of sampling method: With and without SMOTE

3.3.2 Information Gain and Feature Selection

According to Figure 3, “Registration” gained the highest information for “Target”, followed by “Age”, “Channel_Code”, “Vintage”, “Credit_Product”, “Region_Code”, “Occupation_Salaried”, “Occupation_Entrepreneur”, and “Active”. However, the other remained four variables; “Dependent”, “Marital_Status”, “Account_Type”, and “Years_at_Residence” which contributed the lowest information gain for “Target” was removed for the model simplicity.

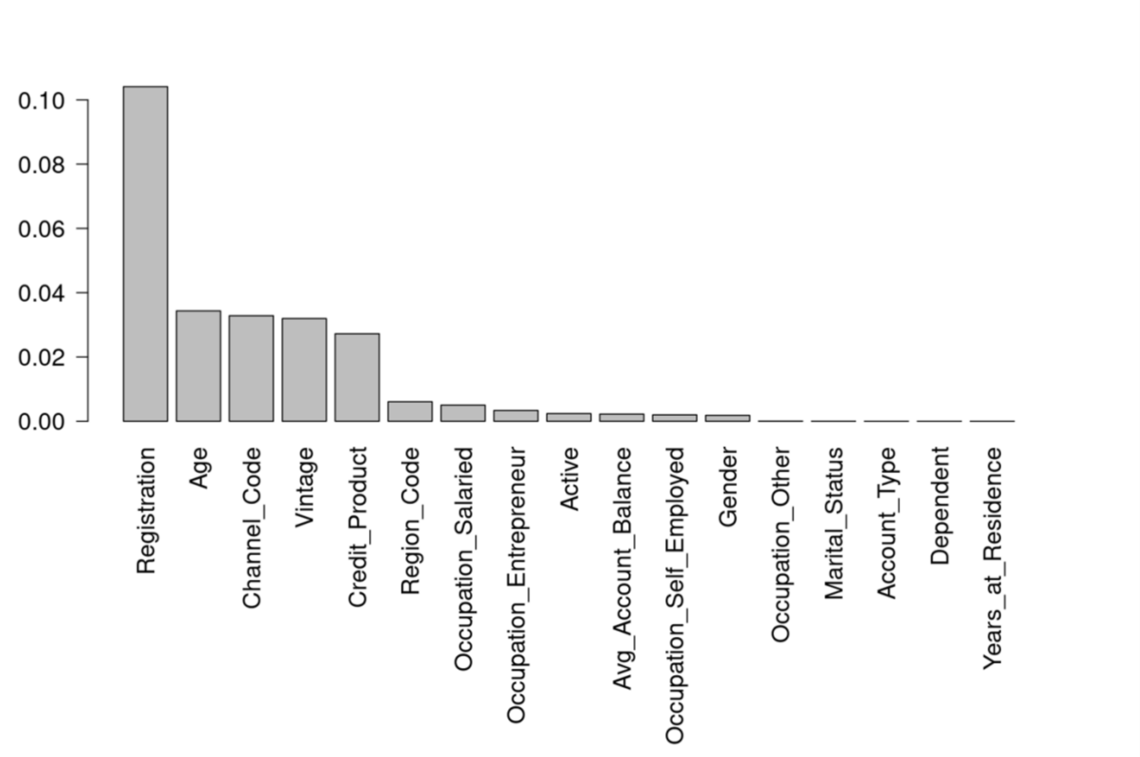


Figure 3: Information gain

3.4 Data Modelling

3.4.1 Introduction

The next step in CRISP-DM is data modelling. The primary goal of data modelling is to identify customers who are inclined to buy the product as well as those who are not. We implemented 5 different classification modelling techniques – Decision Tree, Linear Discriminant Analysis, Logistic Regression, Random Forest, and SVM. After analysing the evaluation metrics, we found that Random Forest model was the best model which was more suitable to the data and predicted with more accuracy.

3.4.2 Model Building

The Random Forest is a classification model which builds individual decision trees and predicts the value based on the class which has the maximum votes. Random Forests are an ensemble learning method for classification that operates by constructing multiple decision trees during training time and outputting the class that is the majority vote of the classes output by individual trees (Liaw and Wiener, 2002).

Modelling approach

We used 4 different approaches for modelling with different number of attributes. Based on information gain and feature selection in the previous section:

- The first model has all attributes from the Information gain chart as they have an impact towards the “Target”. The first model has all attributes from the Information gain chart as they have a clear impact towards the “Target”.
- The second model is built using 3 attributes including “Registration”, “Age” and “Channel_Code” as they have the highest information gain and are highly correlated.
- The third model has 5 attributes, in addition to the first model, “Vintage”, “Credit_Product” as they have visible information gain and are moderately correlated.
- The fourth model has 7 attributes, adding “Region_Code” and “Occupation_Salaried”, to the second model.

The 4 approaches are applied to the 5 modelling algorithms. Finally, 20 models are built and will be evaluated by Accuracy, Precision, Recall, F1 score and AUC in the next section.

3.5 Model Evaluation

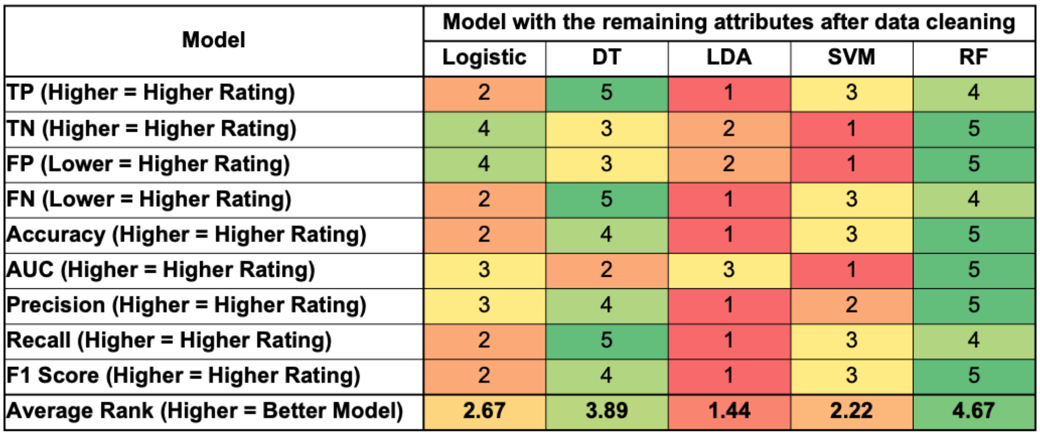
3.5.1 Evaluation Results

We chose the Random Forest model with all remaining attributes after data cleaning due to its consistency across all metrics, as well as its highest precision rate and AUC.

Having more attributes in the model ensures a more balanced trade-off between precision and recall. According to Table 4, there is only a 7% difference, which is the smallest compared to other Random Forest models with fewer attributes. By incorporating more attributes, we can capture a broader range of customer characteristics, thereby enhancing predictive capabilities and customer classification performance. In our case, the mentioned attributes are considered essential information that can be easily obtained from every client. Therefore, it is reasonable to include all these attributes in our model testing.

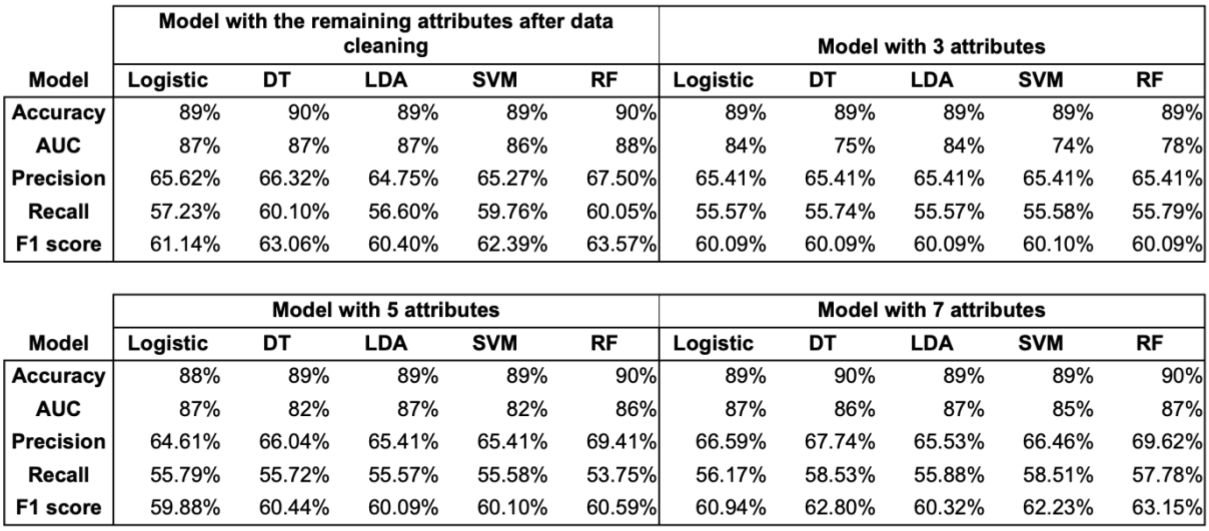
As shown in Table 3, Random Forest demonstrated its pre-eminence in terms of stability as the model obtained the highest rating in most metrics. One of the metrics that achieved the highest is precision, which attained 67.5%, according to Table 4. It implies that the model is the most effective in identifying truly uninterested customers, further supporting our objective of avoiding unnecessary expenses.

Equally important, a higher and more leftward ROC curve indicates better decision or performance testing. It also corresponds to a higher probability of correctly ranking instances, resulting in improved overall classification performance (Metz, 1978). As shown in Figure 4, our model achieved the highest AUC rate of 88%, indicating the best performance in distinguishing between positive and negative instances.



Model	Model with the remaining attributes after data cleaning				
	Logistic	DT	LDA	SVM	RF
TP (Higher = Higher Rating)	2	5	1	3	4
TN (Higher = Higher Rating)	4	3	2	1	5
FP (Lower = Higher Rating)	4	3	2	1	5
FN (Lower = Higher Rating)	2	5	1	3	4
Accuracy (Higher = Higher Rating)	2	4	1	3	5
AUC (Higher = Higher Rating)	3	2	3	1	5
Precision (Higher = Higher Rating)	3	4	1	2	5
Recall (Higher = Higher Rating)	2	5	1	3	4
F1 Score (Higher = Higher Rating)	2	4	1	3	5
Average Rank (Higher = Better Model)	**2.67**	**3.89**	**1.44**	**2.22**	**4.67**

Table 3: Rating of evaluation metrics on different models



	Model with the remaining attributes after data cleaning					Model with 3 attributes				
Model	Logistic	DT	LDA	SVM	RF	Logistic	DT	LDA	SVM	RF
Accuracy	89%	90%	89%	89%	90%	89%	89%	89%	89%	89%
AUC	87%	87%	87%	86%	88%	84%	75%	84%	74%	78%
Precision	65.62%	66.32%	64.75%	65.27%	67.50%	65.41%	65.41%	65.41%	65.41%	65.41%
Recall	57.23%	60.10%	56.60%	59.76%	60.05%	55.57%	55.74%	55.57%	55.58%	55.79%
F1 score	61.14%	63.06%	60.40%	62.39%	63.57%	60.09%	60.09%	60.09%	60.10%	60.09%

	Model with 5 attributes					Model with 7 attributes				
Model	Logistic	DT	LDA	SVM	RF	Logistic	DT	LDA	SVM	RF
Accuracy	88%	89%	89%	89%	90%	89%	90%	89%	89%	90%
AUC	87%	82%	87%	82%	86%	87%	86%	87%	85%	87%
Precision	64.61%	66.04%	65.41%	65.41%	69.41%	66.59%	67.74%	65.53%	66.46%	69.62%
Recall	55.79%	55.72%	55.57%	55.58%	53.75%	56.17%	58.53%	55.88%	58.51%	57.78%
F1 score	59.88%	60.44%	60.09%	60.10%	60.59%	60.94%	62.80%	60.32%	62.23%	63.15%

Table 4: Evaluation results of 20 models

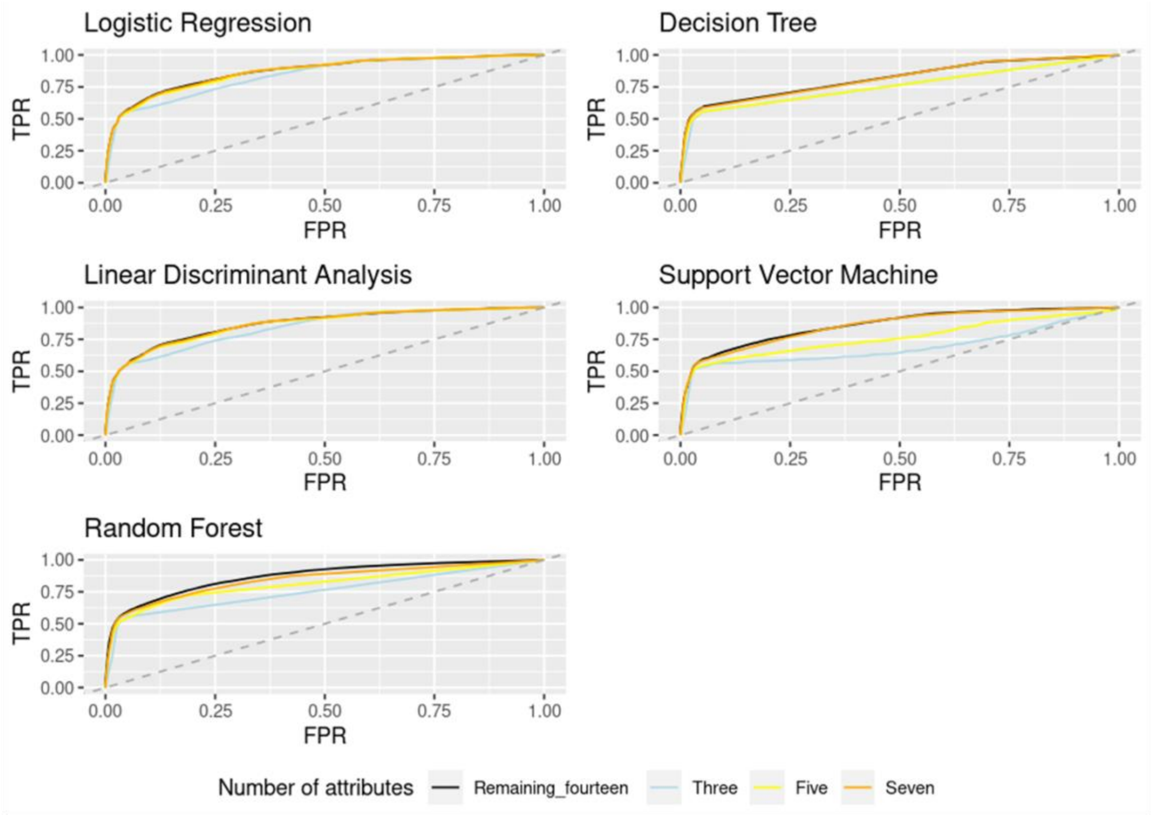


Figure 4: ROC Curves of 20 models

3.5.2 Overfitting

To further enhance our model evaluation, we examined the overfitting of the model. According to Table 5, the accuracy of both training and testing data are similar at around 90%, which indicates there is no sign of overfitting. With high interpretability in unseen data, Random Forest could be seamlessly applied to real-life practice.

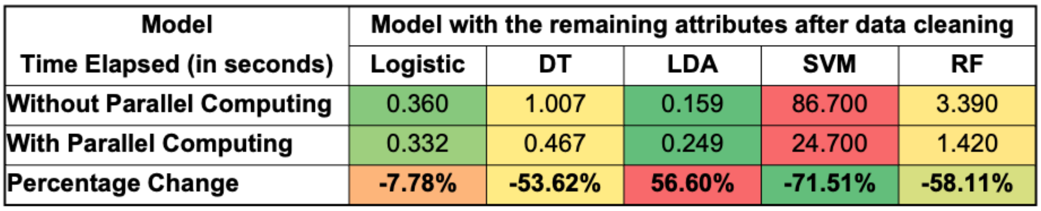
Training Data		Actual	
		Purchase	Did Not Purchase
Prediction	Purchase	14778	5287
	Did Not Purchase	8017	125919
Training Accuracy: 91.4%			

Testing Data		Actual	
		Purchase	Did Not Purchase
Prediction	Purchase	5866	2821
	Did Not Purchase	3902	53410
Testing Accuracy: 89.8%			

Table 5: Confusion matrix and accuracy rate of Random Forest on training and testing data

3.5.3 Computational Efficiency

In real-life application, the ability of interpreting data and making predictions in a short period of time is important. For this reason, parallel computing was applied to split the testing data across 4 CPU cores and predict the results simultaneously. According to the code execution time shown in Table 6, Random Forest significantly dropped from 3.40 to 1.42 seconds. It was also stated that the predicted results with parallel computing had no significant difference between that without parallel computing (Azizah, Riza and Wihardi, 2019). Hence, Random Forest could provide a high data processing speed to cope with large datasets while maintaining model accuracy.



Model	Model with the remaining attributes after data cleaning				
Time Elapsed (in seconds)	Logistic	DT	LDA	SVM	RF
Without Parallel Computing	0.360	1.007	0.159	86.700	3.390
With Parallel Computing	0.332	0.467	0.249	24.700	1.420
Percentage Change	**-7.78%**	**-53.62%**	**56.60%**	**-71.51%**	**-58.11%**

Table 6: Code execution time on testing data with and without parallel computing

4 Conclusion

In conclusion, this report has demonstrated the critical role of data mining through the application of the CRISP-DM framework. By implementing 5 different classification models, we identified the Random Forest model as the most effective approach for World Plus's objective of optimising customer lead identification and marketing expenditures. RF stood out due to its consistency across all evaluation metrics, particularly in terms of Accuracy, Precision and AUC.

To guarantee that the model operates effectively in the future, we think it will be necessary to continuously improve feature selection and parameter tuning to adjust to risks such as changes in consumer behaviour and market trends.

5 Reference Lists

Anand, V. & Mamidi, V. 2020, "Multiple Imputation of Missing Data in Marketing", IEEE, , pp. 1.

Azizah, N., Riza, L.S. & Wihardi, Y. 2019, "Implementation of random forest algorithm with parallel computing in R", Journal of physics. Conference series, vol. 1280, no. 2.

Blagus, R. & Lusa, L. 2013, "SMOTE for high-dimensional class-imbalanced data", BMC bioinformatics, vol. 14, no. 1, pp. 106-106.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

López, V., Fernández, A., García, S., Palade, V. & Herrera, F. 2013, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics", Information sciences, vol. 250, pp. 113-141.

Metz, C.E. 1978, "Basic principles of ROC analysis", Seminars in nuclear medicine, vol. 8, no. 4, pp. 283-298.

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

Saeed, S.E., Hammad, M. & Alqaddoumi, A. 2022, "Predicting Customer's Subscription Response to Bank Telemarketing Campaign Based on Machine learning Algorithms", IEEE, , pp. 1474.

Win, T.T. & Bo, K.S. 2020, "Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm", IEEE, pp. 236.

Wirth, R. and Hipp, J., 2000, April. CRISP-DM: Towards a standard process model for data mining. “In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining”, (Vol. 1, pp. 29-39).