TECHNICAL REPORT

Table of Contents

Introduction

World Plus is a mid-size private bank facing the challenge of accurate lead identification while cross-selling a new term deposit product to their existing customers. Cross-selling has been proven to bring about positive effects on customer lifetime value (Blattberg, Malthouse and Neslin, 2009) and customer loyalty and retention (Dahana et al., 2020). This project aims to develop a robust lead prediction system to maximise the profit earned from cross-selling while minimising marketing expenses. This project follows a CRISP-DM methodology in which multiple machine learning algorithms, including Support Vector Machines (SVM), Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, and Naive Bayes were employed. The models were assessed from both technical perspectives via Confusion Matrix, Receiver Operator Characteristic (“ROC”) graph and business perspectives with an expected profit and loss calculation. Naïve Bayes was selected as the most profitable model for World Plus, with the highest recall rate of 68.21%.

Literature Review

Mohammed et al., in 2019, aimed at identifying the best-performing machine learning algorithms for success prediction in bank telemarketing. The results indicated that Random Forest exhibited the highest accuracy, while Perceptron had the lowest. In terms of precision, Sequential Minimal Optimization performed the best. Due to several common features in the dataset of the paper and this project (demographic information about customers, bank balance and vintage), the modelling and evaluation parts could be applied to analyse the results and interpret what they signify.

Similarly, Madaan et al., in 2021 also employed Random Forest and Decision Tree employed in their research to assess borrower risk and loan feasibility. A detailed exploratory data analysis preceded the modelling phase, shedding light on various relationships within the dataset. The comprehensive evaluation, including confusion matrices and classification reports, indicated that Random Forest was more effective for loan prediction with a higher accuracy rate (80% as compared to 73% achieved by Decision Tree) within the given data set.

Another research using data mining techniques was by Karvana et al., (2019) which attempted to predict customer churn in a private bank in Indonesia. Analysing a dataset with 57 attributes using the CRISP-DM methodology, SVM, with a comparison of 50:50 sampling, performed better than Decision Tree, Neural Network, Naïve Bayes, and Logistic Regression. Karvana et al (2019) and the project shared some common traits, including the same CRISP-DM methodology and some common data attributes such as customer demography and bank balance. Therefore, the research methodology was referenced, especially its model selection method, which was done by comparing the estimated profit and loss generated by each model.

Vafeiadis et al. (2015) also tackled the problem of customer churning prediction by comparing the performance of multi-layer Artificial Neural Networks, Decision Trees, Support Vector Machines, Naïve Bayes, and Logistic Regression, applying boosting techniques which showed a significant improvement in model performance. As recall and precision were also the same key evaluation criteria for this project, the data-boosting technique could be referenced.

Another paper on data mining in the financial sector was Sadikin & Alfiandi (2018) on predicting future customer behaviours. Two data mining techniques, C4.5 and Naive Bayes, were applied in the research, with C4.5 showing the higher accuracy. Additionally, k-fold cross-validation was applied to confirm the higher accuracy of C4.5 as compared to Naive Bayes. This project could refer to the trial of different data splitting ratios and k-fold validation methods owing to certain common customer attributes with the paper.

Moreover, Lim and Singh (2020) addressed the data imbalance in their study to detect SMS spam, which was also a key issue in this project since only 14.8% of customers in the data set bought the new deposit product. The paper contained three machine learning models (C5.0, SVM, Naive Bayes) to analyse their sensitivity and global performance in a class imbalance scenario. AUC, sensibility, and specificity are considered important model evaluation indicators, because specific measures such as precision, recall, and F1 were not suitable. As this study aimed at identifying True Negative whilst True Positive is the goal of this project, the reversal can be inferred, in which precision, recall and F1 should be used as the evaluation criteria.

Business Understanding

Launching a new term deposit product, World Plus Bank needs a lead prediction model to tactically target prospective customers via effective communication channels while minimising sales and marketing expenses. Therefore, this project attempts to build a machine learning algorithm to accurately predict lead conversion to maximise the net profit for World Plus's new product.

CRISP-DM methodology was applied in the project. After analysing data to understand the characteristics of each variable, some data preparation steps would be conducted to mitigate data issues such as imbalance and missing values before splitting it into training and testing sets. Six models were then built from the training data and validated through the test data. Both technical and business evaluation methods would then be applied to select the best model to meet the project's objective.

Data Understanding

220,000 historic customer records were provided by Word Plus with 15 independent variables and one dependent variable, "Target" as illustrated in Table 1.

Table 1 - Table of all variables in the provided data set

Variable Name	Variable Definition	Variable Type
ID	Customer identification number	Categorical
Gender	Female/Male	Categorical
Age	Customer age	Numerical
Dependent	Whether the customer has a dependent or not	Categorical
Marital_Status	Customer's marital state	Categorical
Region_Code	Region code of customer's residence	Categorical
Years_at_Residence	The duration the customer lived in the current residence	Numerical
Occupation	The customer's occupation type	Categorical
Channel_Code	The channel used to reach the customer when they opened their bank account	Categorical
Vintage	How long the customer has stayed with the bank (month)	Numerical
Credit_Product	Whether the customer has any active credit products	Categorical
Avg_Account_Balance	Average account balance for the customer in the last 12 months	Numerical
Account_Type	Account level of the customer	Categorical

Active	If the customer has been active in the last 3 months	Categorical
Registration	Whether the customer visited the bank for the offered product registration	Categorical
Target	If the customer purchased the product	Categorical

Data Preparation

Initially, “ID” variable with little information about customers was removed from the data set. To ensure the proper functioning of machine learning algorithms, binary categorical variables were converted to numerical variables while categorical variables with more than three levels were encoded by one hot encoding for “Marital_Status”, "Occupation", "Channel_Code", "Account_Type" and a target encoding for “Region_Code” with 35 levels (Singh, 2018).

Additionally, the provided data set comprised two main issues, which were (i) imbalanced dependent variable with only 14.8% of total leads converted and (ii) significant portion (8.3%) of missing values for “Credit_Product” variable. To solve (i), Random Over Sampling Examples (“ROSE”) method was applied to the training data, which was stratified partitioned from the total data set at the proportion of 0.7 (Menardi and Torelli, 2012). All three methods (oversampling, under-sampling, and both-sampling) were separately tried in the training data to compare the model performance among themselves as well as the training data without balancing methods. To address (ii), a trial was also conducted, in which a comparison between models built from data with missing values removed, kept untouched and replaced with the mode value was conducted. The results showed that oversampling training data with replaced missing values had the highest performance.

Modelling

With oversampled and encoded training data of 25 variables, six classification models were utilised to predict the lead conversion for the new products, namely, Decision Tree, Support Vector Machine (SVM), Random Forest, Logistic Regression, Gradient Boosting and Naive Bayes, because of their high performance in predicting binary dependent variable, especially in the banking industry (Karvana *et al.*, 2019; Mohammed et al, 2019).

Information gain method was applied to assess which attributes were informative for predicting the “Target” variable (Provost & Fawcett, 2013, p.52) in the training data. However, the removal of attributes with minimal information gain did not significantly improve the model performance, therefore, to avoid overfitting, all attributes remained in the final model version.

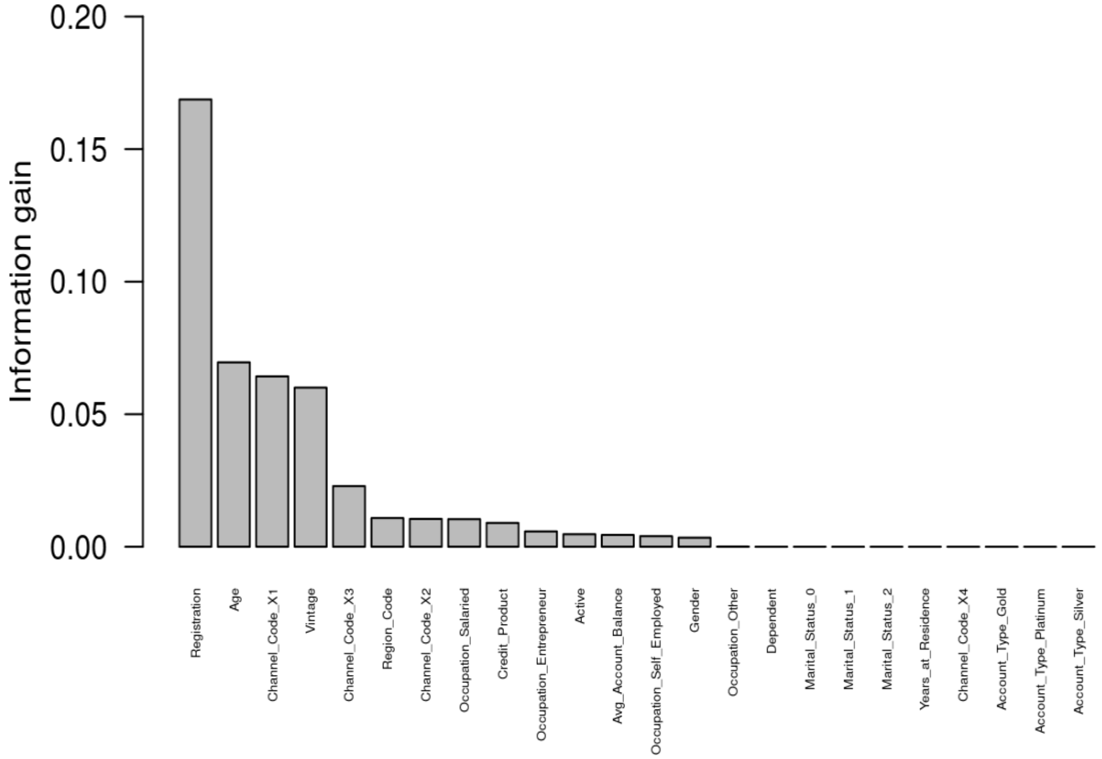

Figure 1 - Information gain of attributes in the training data

Model Evaluation

In the assessment stage, each model generated true positive, false positive, true negative, and false negative, which were mapped into the confusion matrix and area under curve (“AUC”) to obtain their performance as shown in Table 2.

Table 2 - Model performance

Model	Accuracy	Recall	Precision	AUC
Decision Tree	83.74%	65.31%	46.41%	0.8211
SVM	89.69%	59.85%	66.86%	0.8462
Random Forest	**90.38%**	53.99%	**73.82%**	0.8712
Logistic Regression	88.19%	60.06%	60.00%	0.8566
Gradient Boosting	89.19%	61.96%	63.81%	**0.8809**
Naïve Bayes	80.1%	**68.21%**	39.85%	0.8391

Naïve Bayes model had the highest recall rate, which represented the percentage of how many customers the model could accurately predict out of the total customers who had bought the product (Kilinc & Rohrhirsch, 2022). This value is an essential criterion because World Plus wants to successfully turn a potential lead into an actual buyer given the current data set.

Precision is also a crucial metric for consideration, which encapsulates the proportion of accurate predictions of converted customers over the overall number of positive predictions generated by the model (Kilinc & Rohrhirsch, 2022). This ratio can be applicable in calculating the loss incurred by the bank if wrongly predicted leads were followed (Karvana, et al., 2019). The model with the highest precision rate is Random Forest with a rate of 73.82%. However, its recall rate was significantly lower than the other models.

Besides, a ROC graph was also plotted to compare the performance among the models themselves as well as the random baseline model (Provost, F, & Fawcett, T, 2013, p.215). There were no significant visual variances in the model performances on the ROC graph.

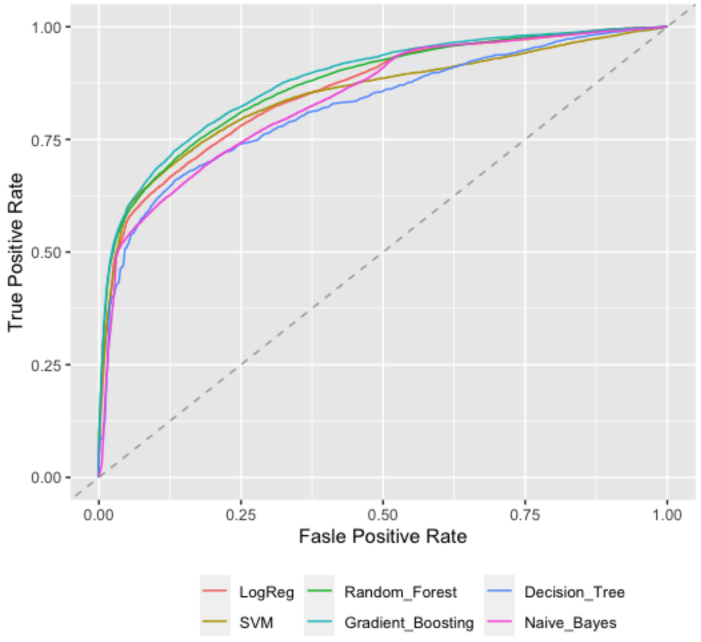

Figure 2 - ROC chart for all models

Since the key objective was to find the model maximising the bank’s profit, an expected value method was applied in the model selection process (Provost & Fawcett, 2013, p.198-203).

As the new product is a term deposit, the revenue obtained per correctly predicted lead was assumed to be the holdable funds, equivalently, the average balance of customers’ bank accounts over the last 12 months (Karvana *et al.*, 2019). The mean bank account balance in the validation data was GBP 1,115,099 per customer.

Another fundamental assumption was the potential loss arising from misguided targeting, which equalled the probability of False Positives times cost per lead, mainly comprising sales and marketing expenses (Karvana *et al.*, 2019). The average cost per lead for the financial services sector is forecasted to be USD 658 in 2024 (Bailyn, 2023), equivalent to GBP 522.

The last assumption was the opportunity costs incurred if the bank mistakenly ignored potential customers. This loss per customer was equivalent to the foregone revenue, which was assumed to be the average balance of customers' bank accounts over the last 12 months.

Based on these assumptions, the estimated net profit per customer for each model was calculated as indicated in Table 3.

Table 3 - Expected profit per customer from the models

Model	TP	FP	TN	FN	Expected profit (GBP)
Decision Tree	6,362	7,347	48,875	3,380	50,295
SVM	5,831	2,890	53,332	3,911	32,383
Random Forest	5,260	1,865	54,357	4,482	13,092
Logistic Regression	5,851	3,901	52,321	3,891	33,052
Gradient Boosting	6,036	3,424	52,798	3,706	39,309
Naïve Bayes	6,645	10,030	46,192	3,097	**59,838**

Conclusion

In conclusion, with demographic, basic financial information and key communication channels of customers, Naïve Bayes model with the highest recall rate (68.21%) and an estimated profit value of GBP 59,812 per customer was selected to be the best lead prediction model for World Plus. Upon being assigned as the consultant, the project team will further improve the model's performance by collecting more data dimensions such as customers' banking habits, the bank's marketing efforts (Kilinc & Rohrhirsch, 2022), customer satisfaction level (Karvana, et al., 2019). The estimation of model's expected profit and loss will also be more accurate if data about customer lifetime value and cost per lead is provided by the bank.

References

Bailyn, E., 2023. FirstPageSage. [Online]
Available at: https://firstpagesage.com/reports/average-cost-per-lead-by-industry/

Blattberg, R. C., Malthouse, E. C. & Neslin, S. A., 2009. Customer Lifetime Value: Empirical Generalizations and Some Conceptual Questions. Journal of Interactive Marketing, 1 5.23(2).

Dahana, W. D., Miwa, Y., Baumann, C. & Morisada, M., 2019. Relative importance of motivation, store patronage, and marketing efforts in driving cross-buying behaviors. Journal of Strategic Marketing, 28 8, 30(5), pp. 481-509.

Karvana, K. G. M., Yazid, S., Syalim, A. & Mursanto, P., 2019. Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. Bali, Indonesia, IEEE.

Kilinc, M. S. & Rohrhirsch, R., 2022. Predicting customers’ cross-buying decisions: a two-stage machine learning approach. Journal of Business Analytics, 9, Volume 6, pp. 1-8.

Madaan, M. et al., 2021. Loan default prediction using decision trees and random forest: A comparative study. Rajpura, India, IOP Publishing Ltd.

Mohammed, A. A., Olalere, M. & Mohammed, A. I., 2019. Comparative Analysis of Performance of Different Machine Learning Algorithms for Prediction of Success of Bank Telemarketing. s.l., Federal University of Technology, Minna Institutional Repository.

Provost, F. & Fawcett, T., 2013. Data. In: Data Science for Business : What You Need to Know about Data Mining and Data-Analytic Thinking. s.l.:O'Reilly Media, pp. 51 - 61, 194-203, 214-222.

Sadikin, M. & Alfiandi, F., 2018. Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk. International Journal of Electrical and Computer Engineering, 8(6).

Singh, N., 2018. [Online]
Available at: https://www.datacamp.com/tutorial/encoding-methodologies [Accessed 1 12 2023].

Vafeiadis, T., Diamantaras, K., Sarigiannidis, G. & Chatzisavvas, K., 2015. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, Volume 55, pp. 1-9.