Contents

1 Introduction

In today's banking landscape, Machine Learning (ML) is crucial in increasing customer engagement and operational efficiency. World Plus, a major player in the financial sector, plans to benefit from ML for its lead conversion strategies, with the goal of predicting customer behaviours for its new term deposit product. This system will optimise targeting across various channels like call centres, live chat, email, and social media, and identify the potential clients who are likely to convert. Thus, this paper aims to build a customer prediction system using the dataset provided by World Plus to enhance lead conversion.

2 Literature Review

Our team evaluate several models of Machine Learning to determine the model to utilize.

2.1 Review on logistics regression

Constantine (2015) describes Logistics Regression as a powerful tool of data analytics that enables decision makers to identify market segments that are more likely to respond to certain marketing actions. Area under the receiver operating curve (AUC) is utilised as one of our team parametric of performance as AUC reflects sensitivity and specificity as individual class performance metrics over all threshold (Gordini and Veglio 2015).

2.2 Predictive ability of classification systems

Charles X. Ling, Jin Huang, and Harry Zhang (2003), in their paper on comparing the predictive ability of classification systems, establish that area under the ROC (Receiver Operating Characteristics) curve, or AUC, provides a better measure than accuracy. It uses discrimination and consistency to establish this result. We also use it in our study/work to find the optimal values of threshold and different parameters.

2.3 Review on SVM

Developed at AT&T Bell Laboratories by Vladimir Vapnik. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. There are numerous hyperplanes that could be used to separate the two classes of data points. Our goal is to find a plane with the greatest margin, or the greatest distance between data points from both classes. Maximising the margin distance provides some reinforcement, allowing future data points to be classified with greater certainty.

2.4 Review on random forest

Xinyu Miao and Haoran Wang (2022) employed three machine learning approaches to forecast customer churn, determining that Random Forest exhibited superior performance evaluated based on ROC & AUC and confusion matrix. Therefore, the utilisation of the Random Forest method in analysing our data for this report is substantiated by ample reasons.

2.5 Forecasting customer attrition in banking system

The research conducted by Singh, Anik, Senapati, et al. (2023) underscores the critical significance of machine learning algorithms in forecasting customer attrition within the banking sector. Among six distinct predictive models analyzed in the study, the Random Forest algorithm emerged as the most efficacious, particularly in terms of key metrics such as sensitivity and accuracy. Furthermore, the study accentuates the Synthetic Minority Over-sampling Technique (SMOTE) as a sophisticated method for addressing challenges associated with imbalanced datasets.

2.6 Review on decision tree

“C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning” by Rutvija Pandya and Jayati Pandya evaluates data mining classifiers ID3, C4.5, and C5.0, highlighting C5.0's superior accuracy and efficiency in decision tree modelling. It demonstrates C5.0's enhanced speed and accuracy in classification. Therefore, C5.0 algorithm is used in this analysis, benefiting from this paper’s implications.

3 Data Preparation

3.1 Data interpretation

The dataset comprises 220,000 observations with 16 variables. *Target* variable is the focus for model prediction, indicating subscription outcomes. Initial data exploration reveals substantial imbalance in the *Target* column, with only 16.8% of customers subscribing. This result indicates that adjustment on the training data is necessary to ensure an unbiased model. String variables like *Region_Code, Occupation, Channel_Code,* and *Account_Type* required conversion to binary or factor formats for effective data training. *Gender, Credit_Product,* and *Active* were transformed into binary categories.

3.2 Data cleaning

Further analysis showed a skewed distribution in the *Target* variable (85%:15%) and minor missing values in *Credit_Product* (0.083%), but no significant outliers. To prevent overfitting, data cleaning and balancing were performed. missForest() function was used to impute missing values in *Credit_Product*, given its strong correlation with the *Target* variable. The information.gain() function was utilised to filter out irrelevant variables, including *Marital_Status, Years_at_Residence,* and *ID*.

Initial tests showed overfitting with near-perfect accuracy, leading to the decision of performing data cleaning for a more balanced dataset, enhancing model training efficacy. To achieve a balanced dataset, various sampling methods were applied: undersampling, oversampling, bothsampling, and SMOTE. These methods were implemented after removing string values and applying appropriate encoding methods like one-hot, binary, Weight of Evidence. Overall, these steps ensured more effective and unbiased model training.

4 Model Implementation

During the model training phase, the confusionMatrix() function is frequently utilized to evaluate the performance of each model, with the specific notations and descriptions employed in this study detailed in Table 1. Additionally, to determine the most effective approach, all input training data are balanced using four distinct sampling methods in conjunction with two encoding techniques. Remarkably, the combination of SMOTE function and one- hot coding outperforms other methods across all models except for logistic regression model(refer to example results in Figure 1), substantiating its efficacy as an advanced technique for managing unbalanced data, as corroborated by Singh, Anik, Senapati, et al. (2023).

Figure 1: Example results

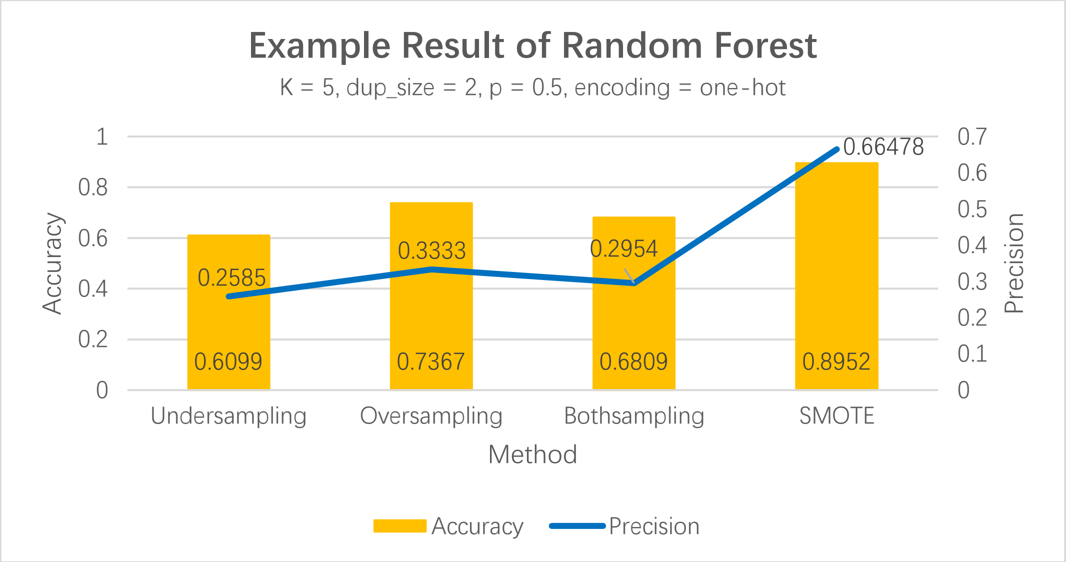


Table 1 Notions and descriptions

Notions	Descriptions
Precision	True Positive / (True Positive + False Positive)
Recall	True Positive / (True Positive + False Negative)
Specificity	True Negative / (True Negative + False Positive)
F1 Score	(2 * Precision * Recall) / (Precision + Recall)
Balanced Accuracy	(Recall + Specificity) / 2
AUC	Area Under the ROC Curve
ROC	Receiver Operating Characteristic Curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
Accuracy	(TP + TN) / (TP + TN + FP + FN)
P	The proportion of training / test dataset
K	The number of nearest neighbors during sampling process
Dup_size	The number or vector representing the desired times of synthetic minority instances over the original number of majority instances
SVM	Support Vector Machine
SMOTE	Synthetic Minority Oversampling Technique

4.1 Decision tree

Decision tree model is a supervised learning method used for classification, predicting the target variable's value based on decision rules derived from data attributes. The model is built with C5.0 algorithm which uses information gain to fit a classification tree.

The decision tree model demonstrated high accuracy, ranging from 88.5% to 89.4% across thresholds of 0.5 to 0.8. Precision varied from approximately 0.59 to 0.66, while recall ranged from 0.58 to 0.63, and F1-scores stayed around 0.60 to 0.62.AUC values varied between 0.82 and 0.85, highlighting significant model performance but also indicating different patterns from accuracy across varying parameters. Analysis of the dataset showcased the impact of varying parameters (p, k, and dup_size) on model's performance. Thresholds around 0.6-0.7 favoured precision, lower thresholds improved recall, and SMOTE usage generally enhanced positive instance capture, aiding in customer prediction.

Experimenting with SMOTE sampling revealed diverse outcomes. Firstly, the best precision and accuracy were observed with 'p' set at 0.7. Secondly, adjusting 'dup_size' while keeping K constant showed 'dup_size' of 2 as optimal in terms of precision and accuracy**.** Thirdly, incrementally increasing K revealed the value of 6 to be most effective.. Therefore, setting the SMOTE parameters as p= 0.7, K= 6, and dup_size= 2, yielded the best results for the decision tree model, achieving an accuracy of 0.8908 and a precision of 0.6357.

4.2 Logistic regression

The Logistic Regression Model is used for analyzing behaviours in dichotomous variables (Constantin 2015). Constantin explains that the model operates on a logit-based calculation, estimating the likelihood of specific events occurring. In this case, the model will be evaluated in a multi-probability acceptance level from 0.9 to 0.1.

The model was evaluated through training with adjustment on five parameters: encoding method, threshold, K value, dup_size, and probability acceptance level. Utilisation of WOE encoding yields higher accuracy and precision than one-hot encoding. Using WOE encoding, a gradual increase in these metrics was observed with increasing split values, peaking at 0.6. Beyond this point, both accuracy and precision declined.

Holding encoding and thereshold constant, k value yields the highest combination of accuracy and precision when set to 8. Similarly, with a constant K value, a dup_size of 2 yields the best results. Adjusting probability acceptance level while holding other parameters constant shows that a probability acceptance level of 0.9 maximized accuracy and precision.

Combinations of encoding method = WOE, p= 0.6, K= 8, dup_size= 2, and probability acceptance level=0.9 in training Logistic Regression model result in highest mix of accuracy and precision valued at 0.8673 and 0.6513, respectively.

4.3 SVM

The SVM algorithm seeks the best hyperplane in an N-dimensional space for separating data points into different classes in the feature space. The hyperplane attempts to maximise the margin between the closest points of different classes. The size of the hyperplane is determined by the number of features. If there are only two input features, the hyperplane is simply a line. If the number of input features is three, the hyperplane becomes a two-dimensional plane. When the number of features exceeds three, it becomes difficult to imagine. Combinations of encoding method = One-Hot, p= 0.6, K= 8, and dup_size= 2 in training SVM model result in highest mix of accuracy and precision valued at 0.8467 and 0.27147, respectively.

4.4 Random forest

In developing the Random Forest model, a comprehensive function was crafted to encompass all aspects of the modeling process, including data encoding, cleaning, model building, and returning results. The randomForest() function is chiefly employed in the model construction phase to formulate the primary model. Outcomes across all trained models show that the Random Forest model consistently achieves high levels of accuracy, AUC, and balanced accuracy.

The thresholds for split data were tested from 0.5 to 0.8, with K ranging from 5 to 9, and dup_size varying between 2 to 4. This was done considering the proportion of the Target variable, which fluctuated from nearly 65%: 34% to a more balanced 49%: 51%. The model's accuracy marginally oscillated between 0.894 and 0.895, precision ranged from 0.651 to 0.664, and recall was noted to be between 0.581 and 0.589. F1 scores exhibited minimal variation, altering by no more than 0.01 from 0.61. Balanced accuracy varied slightly from 0.764 to 0.766, while the AUC was recorded to fall between 0.838 and 0.846.

Figure 2 illustrates that a p-value of 0.8 exhibits the best performance. Consequently, with p at 0.8, a comparative analysis of dup_size values, shown in Figure 3, indicates that a dup_size of 2 is superior. Subsequently, further evaluation with different K values, keeping p at 0.8 and dup_size at 2, is presented in Figure 4, which reveals that a K of 5 gives most optimal outcomes. Consequently, the optimal Random Forest model configuration, with p at 0.8, K at 5, and dup_size at 2, results in an accuracy of 89.52% and a precision of 66.48%.

Figure 2: Comparison of P values

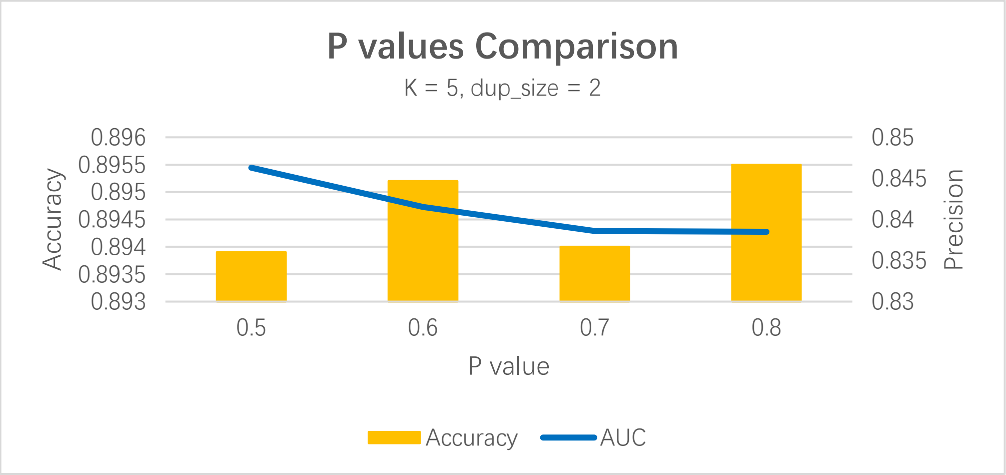


Figure 3: Comparison of dup_size value

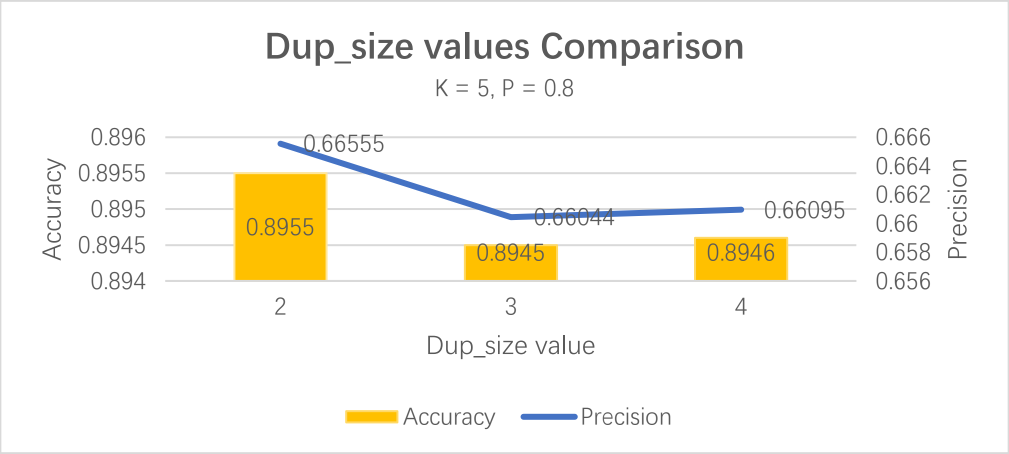
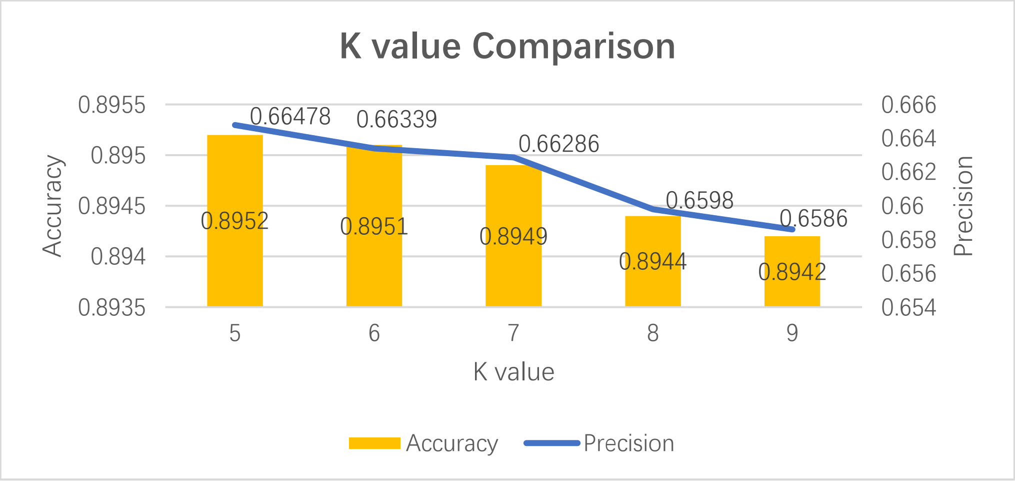


Figure 4: Comparison of K value

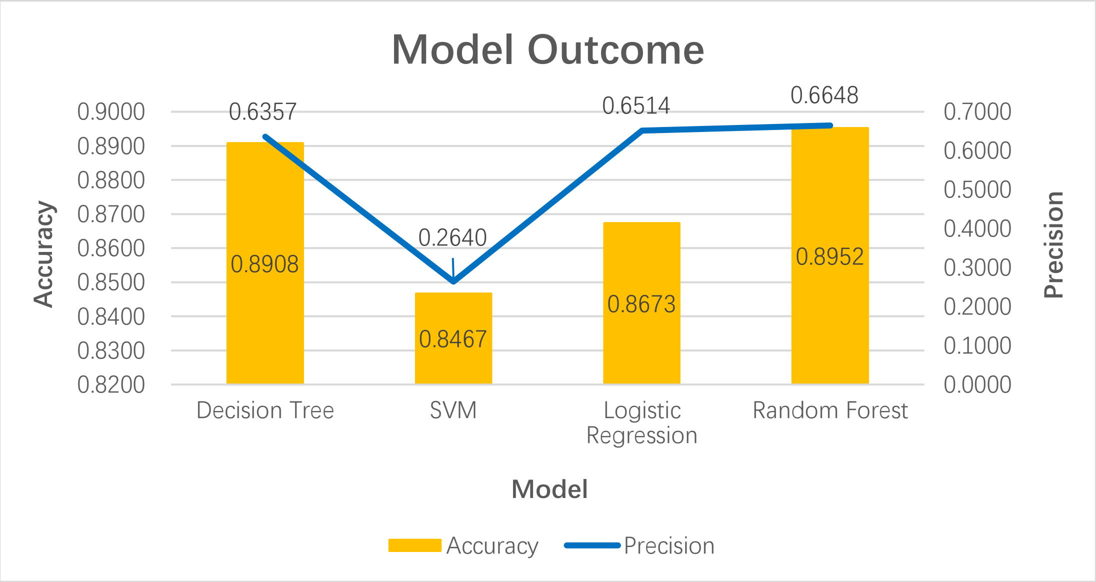


5 Model Evaluation

As posited by Pahul (2023), recall and accuracy are paramount metrics in addressing banking customer attrition, with recall being particularly crucial in accurately predicting customers likely to leave. Given this background, our study has opted to employ accuracy and precision as primary indicators of model performance. Precision is especially significant, as it pertains to the correct prediction of customers who will subscribe to a product — a factor deemed more valuable than identifying departing customers. Furthermore, integrating accuracy with precision ensures the optimization of marketing budgets, effectively preventing wasteful expenditure on advertising to customers who demonstrate no interest in, and ultimately will not purchase, the product.

Figure 5 demonstrates the comparison of the models in terms of Accuracy and Precision, highlighting that Random Forest has the best outcomes for these performance metrics, with an Accuracy of 0.8952 and Precision of 0.6648.

Figure 5: Model Outcome



6 Conclusion

In this report, various ML models were implemented to optimize lead conversion strategies for World Plus, focusing on predicting customer engagement with a new term deposit product. The analysis demonstrates that Random Forest model is the most effective among these techniques, outperforming other models like Decision Tree, Logistic Regression and SVM, with its Accuracy and Precision values of 89.52% and 66.48% respectively. Additionally, the analysis highlights the importance of feature selection and the effectiveness of SMOTE in handling imbalanced datasets. Overall, this report indicates a clear direction for World Plus to apply ML in customer behavior prediction and achieve better lead conversion for its term deposit service. Further innovative strategies could be explored to enhance customer engagement, boost conversion rates, and increase business growth.

References

[1] Ling, Charles & Huang, Jin & Zhang, Harry. (2003). ‘AUC: A Better Measure than Accuracy in Comparing Learning Algorithms’. Canadian Conference on AI. pp. 329-341.

[2] Pandya, Rutvija and Pandya, Jayati. (2015). ‘C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning’. International Journal of Computer Applications. 117(16), pp. 18-21. 10.5120/20639-3318.

[3] Miao, X., and Wang, H. (2022), ‘Customer Churn Prediction on Credit Card Services using Random Forest Method’, Advances in Economics, Business and Management Research, vol. 211, pp. 649-656.

[4] Constantin, C. (2015). ‘Using the Logistic Regression model in supporting decisions of establishing marketing strategies’. Bulletin of the Transilvania University of Braşov Series V: Economic Sciences, 8(57), pp. 43-50.

[5] Singh, P.P., Anik, F.I., Senapati, R., Sinha, A., Sakib, N. and Hossain, E. (2023). ‘Investigating customer churn in banking: A machine learning approach and visualization app for data science and management’. Data Science and Management. DOI: https://doi.org/10.1016/j.dsm.2023.09.002.

[6] Farishy, R. (2023). ‘The Use of Artificial Intelligence in Banking Industry. Master of Management’. Airlangga University, Indonesia. DOI: https://doi.org/10.46799/ijssr.v3i7.447.