Masters Programmes: Assignment Cover Sheet

Student Number:	**5534286, 5550544, 5559314, 5567131, 5578678, 5582906**
Module Code:	**IB9BW0**
Module Title:	**Analytics in Practice**
Submission Deadline:	**6th Dec 2023**
Date Submitted:	**5th Dec 2023**
Word Count:	**1987**
Number of Pages:	**14**
Question Attempted: *(question number/title, or description of assignment)*	**Developing a Lead Prediction System for World Plus**
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 CONTENT

1. Introduction

In today's competitive market, customer acquisition often involves luring them from rivals, but this can be resource intensive. Alternatively, cross-selling to existing customers is a cost-effective strategy, especially for banks, facilitated by machine learning. According to Boustani et al. (2023), focusing on cross-selling to existing customers is a more cost-effective strategy for increasing sales. This report focuses on identifying potential customers for World Plus likely to purchase a new term deposit product, aiming for cost-effective lead conversion.

A Literature Review to contextualize lead prediction's significance was used to start this report. Following that, the Materials and Methods section was used to outline the approach. In the results section, machine learning algorithms were compared and were later discussed in part four.

2. Literature Review

In recent times, data has become a critical asset for companies aiming to comprehend customer buying trends and adjust marketing strategies accordingly. According to Kwiatkowska (2018), utilizing data containing details about consumer purchasing behavior could substantially enhance prediction accuracy. Data mining involves collecting, preprocessing, and modeling data to recognize patterns and correlations among variables for informed decision-making aimed at cost reduction.

Real-life situations often require handling missing data, as noted by Donders et al. (2006). Multiple Imputation by Chained Equations (MICE) is a statistical technique for balancing bias and precision. It creates and analyzes multiple datasets, providing robust estimates and correcting errors compared to simpler methods (Buuren and Groothuis, 2011). Our data mining approach aligns with these insights, therefore MICE techniques were incorporated in our data mining.

According to Moeyersoms and Martens (2015), transformation methods like the dummy method and semantic grouping could result in dimensional challenges when dealing with multi-class variables. Meanwhile, Weight of Evidence (WOE) produced the highest AUC scores while dealing with those variables. Therefore, WOE was implemented in our data mining. The research also recommends conducting WOE on the training dataset to avoid overfitting data.

Geng et al. (2018) proposed that KNN and SVM trained with datasets that went through data balancing methods performed better as compared to those who didn't. Therefore, four common balancing methods, over-sampling, under-sampling, both-sampling, and Synthetic Minority Over-sampling Technique (SMOTE), were used to handle the imbalanced dataset.

Sekeroglu's study (2021) used seven models including K-nearest neighbors (KNN), Logistic Regression, Random Forest, and Support Vector Machine (SVM) to focus on accuracy and F1 score. Random Forest (RF) exhibited higher accuracy and F1 scores while Logistic Regression performed best in terms of AUC. Lalwani et al.'s study (2021) demonstrated that XGBoost achieved 80.8% accuracy outperforming logistic regression, naive Bayes, support vector machine, random forest, decision tree, and KNN classifiers with an AUC of 84%. As different sources recommend different models, Decision trees, KNN, RF, Logistic Regression, SVM, and XGBoost were compared to identify the best model for World Plus based on their interpretability, flexibility, binary classification suitability, complexity, and robustness.

3. Materials and Methods

We use the Cross-Industry Standard Process for Data Mining (CRISP-DM; Shearer, 2000) framework, illustrated in Figure 1.

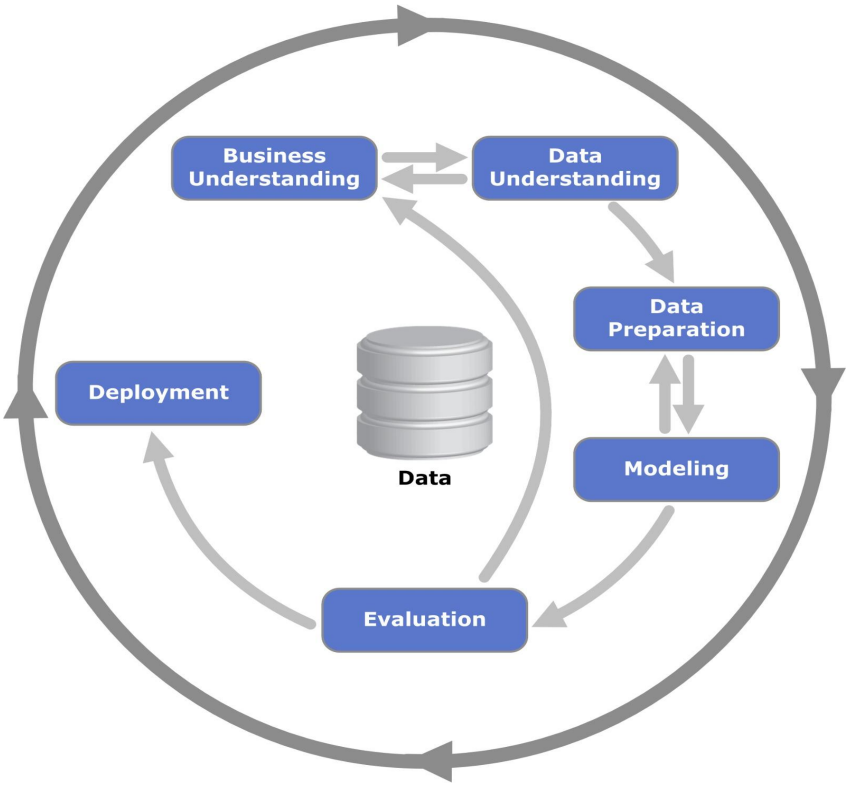

Figure 1: CRISP-DM Framework

3.1. Business understanding

World Plus is facing challenges in identifying potential leads among its customer base, therefore they aim to implement a lead prediction system for its upcoming term deposit product. The main goal is to effectively identify prospective customers likely to commit, optimize communication channels, and reduce unnecessary costs.

3.2. Data understanding

Exploratory data analysis has been conducted to explore the distribution of attributes and detect anomalies. Necessary data pre-processing steps have been implemented to gain a better understanding of the dataset.

3.2.1 Dataset Features

The attributes of the dataset are shown in Table 1.

Table 1: Data Dictionary

St.no	Variable Name	Description
1	ID	customer identification number
2	Gender	gender of the customer
3	Age	age of the customer in years
4	Dependent	whether the customer has a dependent or not
5	Marital_Status	marital state (1=married, 2=single, 0 = others)
6	Region_Code	code of the region for the customer
7	Years_at_Residence	the duration in the current residence (in years)
8	Occupation	occupation type of the customer
9	Channel_Code	acquisition channel code used to reach the customer when they opened their bank account
10	Vintage	the number of months that the customer has been associated with the company.
11	Credit_Product:	if the customer has any active credit product (home loan, personal loan, credit card etc.)
12	Avg_Account_Balance	averge account balance for the customer in last 12 months
13	Account_Type:	account type of the customer with categories Silver, Gold and Platinum
14	Active	if the customer is active in last 3 months
15	Registration	whether the customer has visited the bank for the offered product registration (1 = yes; 0 = no)
16	Target	whether the customer has purchased the product, 0 : customer did not purchase the product, 1 : customer purchased the product

3.2.2 Continuous Variable Analysis

The pair plot, correlation coefficients, and histograms were used to analyze the distribution and relationship of the integer continuous variables in the Age, Avg_Account_Balance, and Vintage attributes as shown in Figure 2.

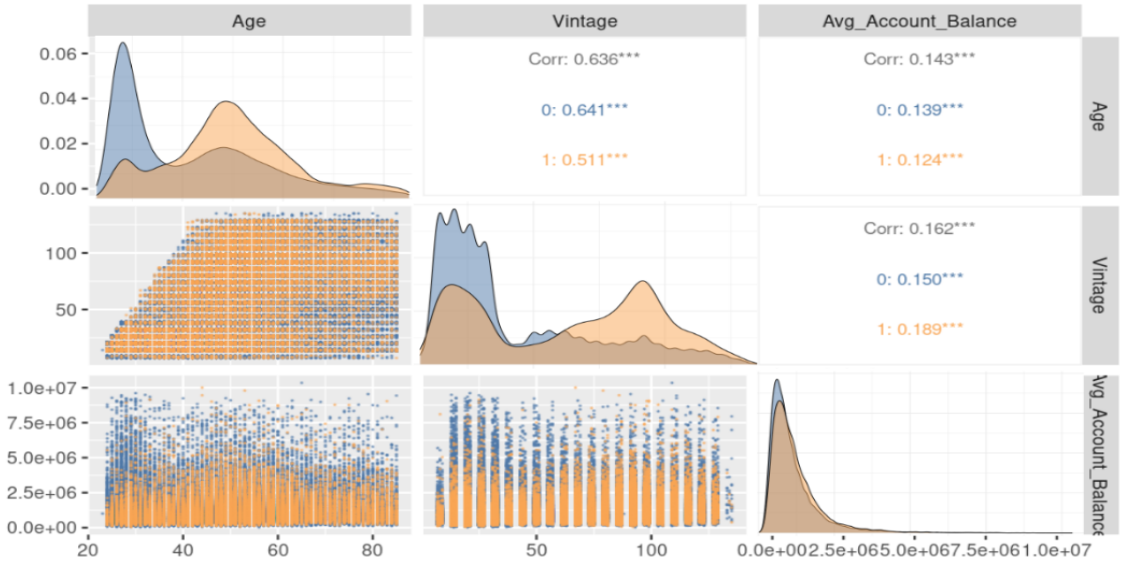

Figure 2: Continuous Variables Pair Plot

The variables showed a strong positive association, especially Age and Vintage, which were highly correlated with class 1. However, Avg_Account_Balance showed a mild positive correlation with Age and Vintage. People over the age of 40 were more likely to respond positively to the previous product. Moreover, the vintage distribution suggested that customers engaged with the bank for over 5 years were more inclined to make purchases. Figure 3 presents a correlation matrix displaying positive relationships among continuous features and emphasizes vintage as having the most significant correlation with our target variable.

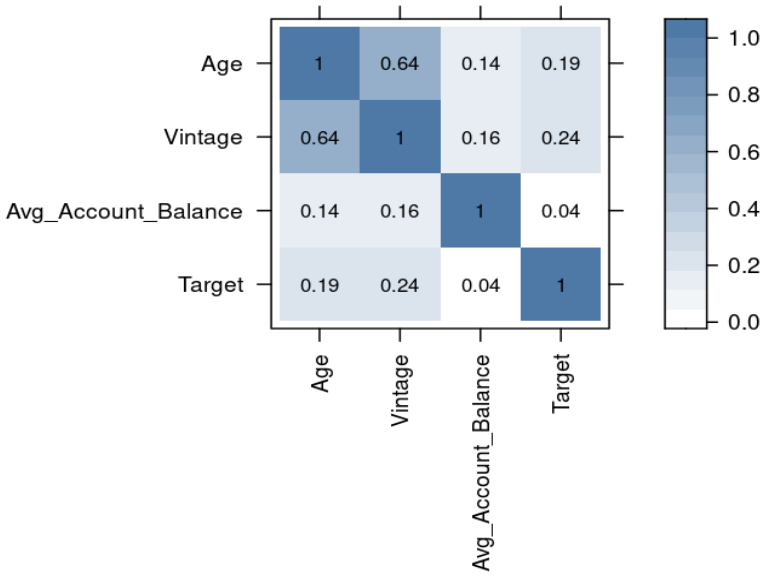

Figure 3: Correlation Heatmap

3.2.3 Categorical Variable Analysis

11 discrete variables were mentioned in the database: Gender, Dependent, Marital_Status, Region_Code, Years_at_Residence, Occupation, Channel_Code, Credit_Product, Account_Type, Active and Registration.

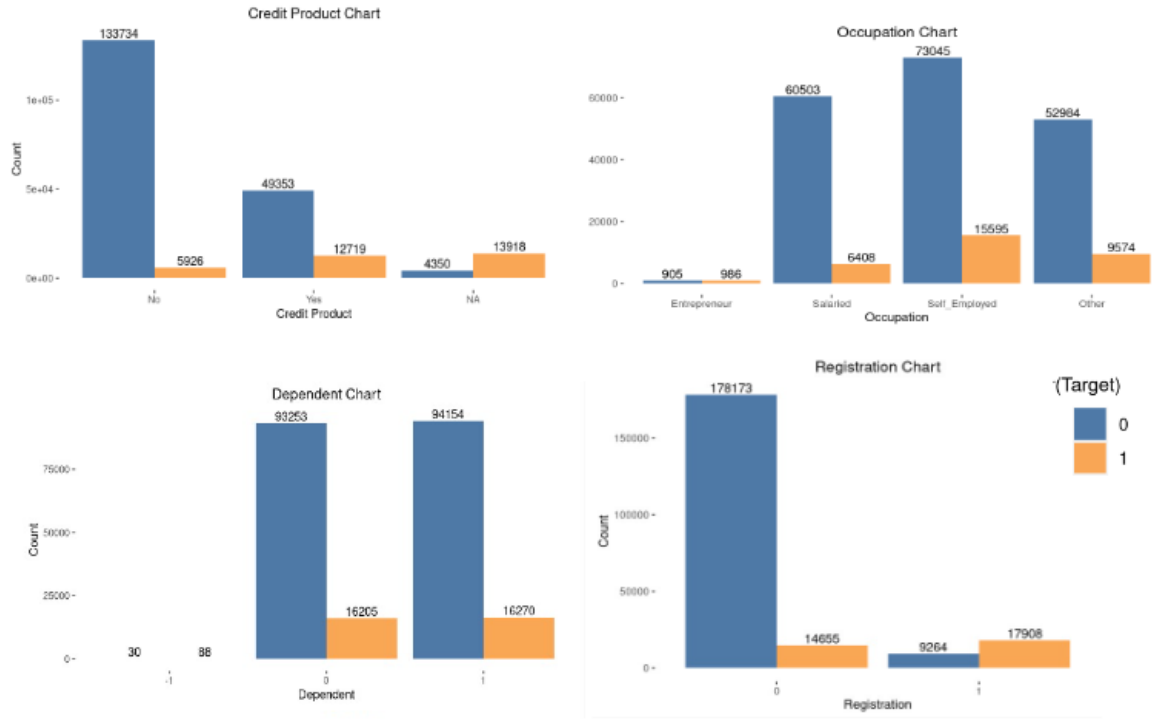

Figure 4: Categorical Variable Analysis Chart

The categorical variables were expressed as bar charts to understand value distributions. Through data visualization, a -1 value was recorded under dependence, which could be seen as an error. Another key attribute recorded was that self-employed was the highest occupation who used the previous product.

Additionally, there were a considerable number of missing values in the credit product type data, with 8.3% recorded as missing. Simply deleting these NA values would not be feasible due to their significant quantity. Ahmad et al. (2019, p.14) attempted to address these missing values by eliminating all features that contained at least one null value. However, this method yielded unsatisfactory results. In the registration data, a larger proportion of individuals who visited the bank purchased the previous product. Moreover, there were cases where purchases were made by individuals who did not visit the bank at all.

3.3. Data Pre-processing

Data pre-processing includes various steps to clean, transform, and organize raw data to improve its quality and suitability for analysis.

3.3.1 Data Transformation

In the preliminary data analysis, no duplicate entries were found, and the ID column was excluded as it did not affect the target variable. Moreover, we adjusted the dependent variable by replacing values coded with -1 to 1 assuming errors in data input. Additionally, there were missing values in the Credit_Product feature which were considered to be missing at random (potentially due to customer preference). Therefore, for handling these missing values in Credit_Product, the 'mice' package was utilized to implement a multivariate imputation technique using logistic regression.

3.3.1.1 Encoding

Label encoding, one-hot encoding, and WOE encoding were utilized to effectively represent categorical variables. Label encoding assigned a unique numerical label to each category in Gender, Credit_Product, Account_Type, and Active. Meanwhile, one-hot encoding created binary columns for each category in Channel_Code and Occupation. Finally, due to the high cardinality of Region_Code, WOE encoding was used to transform categorical data into continuous variables while incorporating information about their relationship with the target variable.

3.3.1.2 Min-Max Scaling

Scaling is crucial for machine learning algorithms that are sensitive to the scale of input features, facilitating equitable and efficient model training. The Age, Vintage, and Avg_Account_Balance attributes from the complete_data dataset were selected for scaling to transform numeric values into a range between 0 and 1.

3.3.1.3 Train-Test Split

In the research, a 70-30 split for training and testing was implemented to ensure sufficient data for model training while also reserving a significant portion for thorough testing. This partition was essential for assessing the model's performance on new, unseen data, which was pivotal in determining its ability to generalize (Raschka and Mirjalili, 2019).

3.3.1.4 Data Balancing

In the classification problem, achieving a balanced output feature was crucial for reliable predictions. The target variable from complete_data, held a significant class imbalance, with 85% in class 0 and only 15% in class 1 as shown in Figure 5.

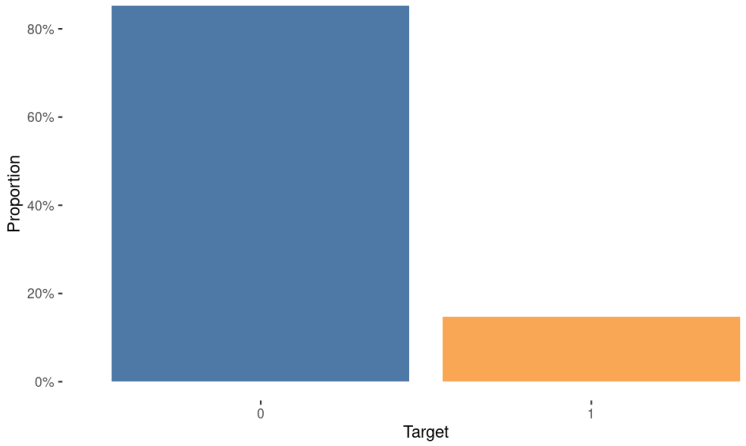

Figure 5: Proportion of Target Variable

Although machine learning techniques were employed to improve accuracy by minimizing errors, not all models consider class balance, potentially leading to poor results. To tackle this challenge, different sampling methods such as over-sampling, under-sampling, both-sampling, and SMOTE were used to evaluate the performance of each model. Furthermore, SMOTE collaborates with the KNN algorithm to generate synthetic data for the minority class (Sekeroglu, 2021, n.p.).

3.4. Feature Selection

The primary aim of conducting feature selection was to identify the features that accurately represent the dataset. Having an excessive number of columns could lead to poor generalization in models (overfitting). Information gain was employed for this purpose, and the outcome was illustrated in Figure 6. The feature values in the table of information gain were all above 0, so all of the columns were utilized.

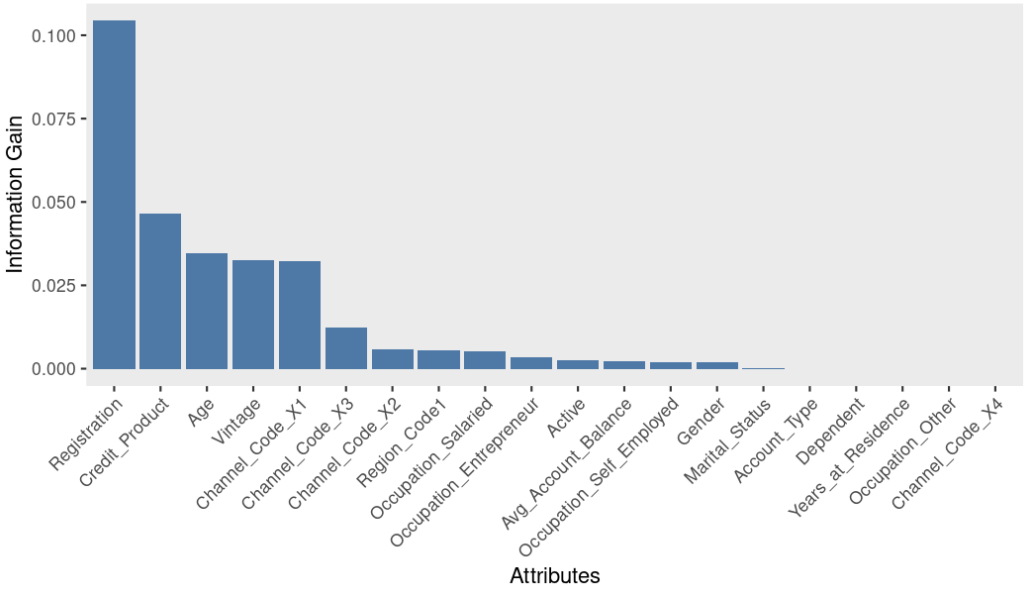

Figure 6: Information Gain Chart

3.5. Modeling

Upon finishing the essential data preprocessing steps, various classification techniques including Decision Tree, K-Nearest Neighbors, Logistic Regression, Random Forest, Support Vector Machine, and XGBoost were employed to detect prospective customers for cross-selling campaigns. Notably, stratified sampling was applied to the KNN model due to its limitation in handling numerous k-values. The effectiveness of these models was then evaluated using metrics such as Area Under the curve, precision, recall, F1 score, and accuracy. The AUC was identified as the most critical metric since it assesses the model across various decision thresholds and is easily interpretable. Precision is also a crucial metric for World Plus because it directly relates to the true positives (correctly identified leads) among all predicted positives. Additionally, precision aids in identifying potential prospects who are more likely to convert. Therefore, precision plays a pivotal role in cost efficiency by reducing resources spent on targeting less promising leads.

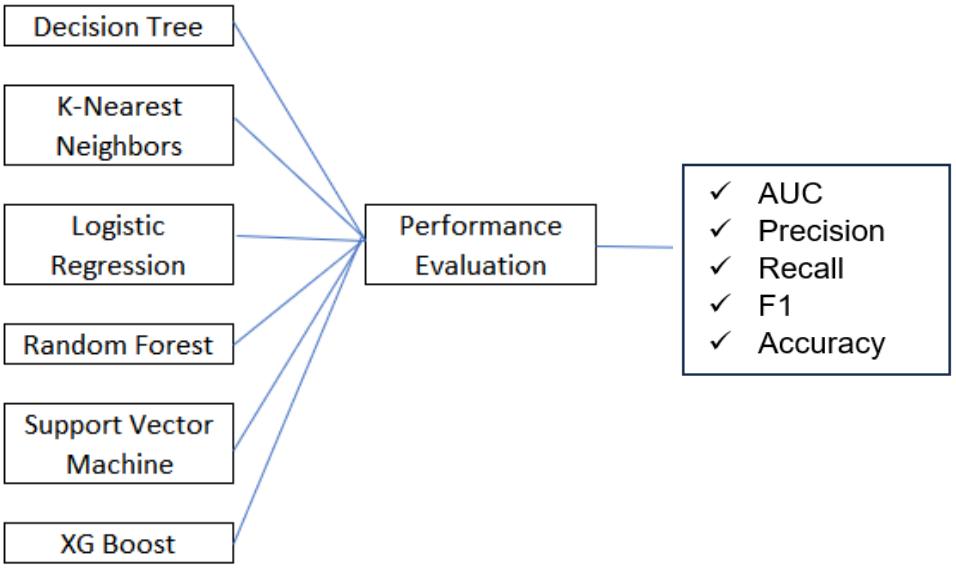

Figure 7: Modelling Framework

4. Results

In our research project, various machine learning models were tested, including SVM, Logistic Regression, Decision Tree, Random Forest, KNN, and XGBoost. The main goal was to identify potential customers for a cross-selling campaign. Moreover, we completed tuning the Random Forest model due to the promising overall performance compared to others and its superior computational efficiency over XGBoost. The tuning results were better at precision, F1, and accuracy.

Table 2: Model Performance Based on Balancing Method

No	Balancing Method	DT	KNN	LR	RF	RF Tuning	SVM	XGBoost
1	both(p=0,5)	0,846	0,941	0,885	0,893	0,887	0,881	0,881
2	over(p=0,5)	0,846	0,946	0,885	0,893	0,880	0,881	0,880
3	smote(2,2)	0,852	0,902	0,884	0,892	0,889	0,872	0,886
4	smote(2,5,1,5)	0,846	0,882	0,884	0,892	0,882	0,877	0,885
5	under(p=0,5)	0,846	0,932	0,885	0,893	0,885	0,880	0,882

Among various balancing methods, oversampling with a probability of 0.5 (p=0.5) produced the most favorable outcome. Therefore, oversampling was adopted for model comparison.

Table 3: Model Performance based on Oversampling

No	Model Name	Balancing Method	Accuracy	Precision	Recall	F1	AUC
1	KNN	over(p=0,5)	0,976	0,934	0,903	0,918	0,946
2	RF	over(p=0,5)	0,852	0,500	0,761	0,603	0,893
3	LR	over(p=0,5)	0,820	0,439	0,786	0,563	0,885
4	SVM	over(p=0,5)	0,800	0,413	0,829	0,551	0,881
5	RF tuning(100)	over(p=0,5)	0,902	0,717	0,562	0,630	0,880
6	XGBoost	over(p=0,5)	0,871	0,550	0,681	0,609	0,880
7	DT	over(p=0,5)	0,775	0,382	0,838	0,525	0,846

The AUC indicates that KNN is the optimal model at 94.6%, followed by Random Forest at 89.3%, and Logistic Regression at 88.5%. Regarding precision, KNN remained the top-performing model at 93.4%, while RF after fine-tuning with the number of decision trees (ntree=100) scored at 71.7%, and XGBoost achieved 55%. The remaining performance can be seen in Figures 8 and 9.

Figure 8: Model Evaluation Metrics

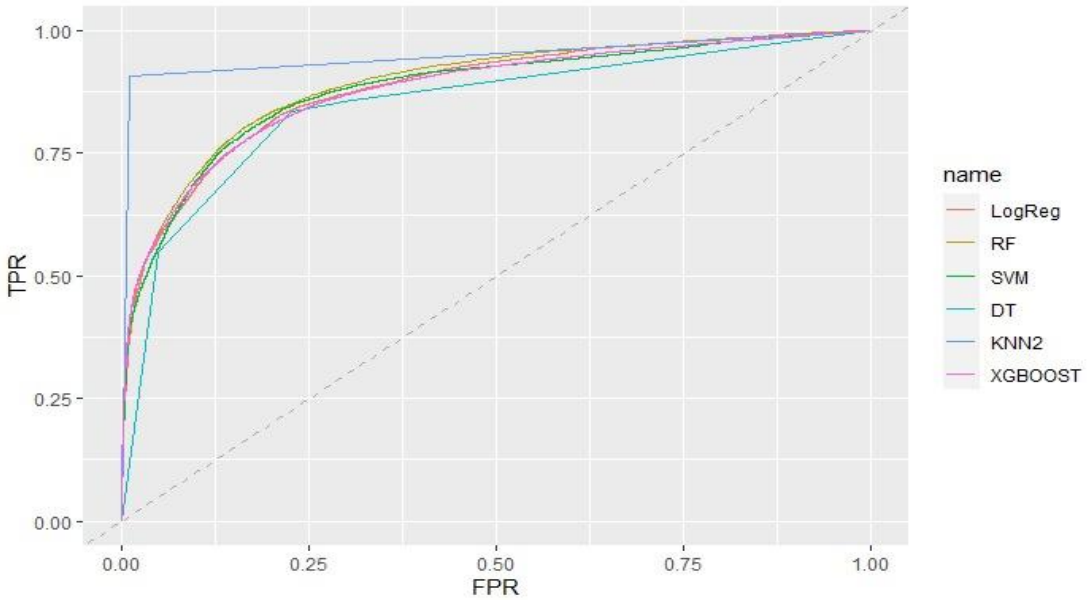

Figure 9: ROC curve for each model

The assessment highlighted the superior performance of the KNN model, not only in terms of accuracy, precision, and recall but also in its outstanding AUC value. This indicated a remarkable ability to distinguish between positive and negative instances compared to the Random Forest and Logistic Regression models. The unique capability of KNN as a case-based learning model was contingent on the k-value employed. In this study, the k-value derived was utilized from the square root of total records in the dataset.

5. Conclusions

To evaluate the efficacy of different machine learning models in categorizing the suitability of cross-selling marketing campaigns, several preprocessing techniques were used, including standardization, handling missing values, data exploration, and encoding. The findings indicate that individuals aged between 40-50 years old exhibited a higher propensity to make a purchase. Furthermore, customers who have been with the bank for 5 or more years and visit during promotional offers demonstrate an increased likelihood of making a purchase. The report conducted a comparative analysis using various machine learning algorithms to minimize unnecessary marketing expenses for World Plus. With only 15% of the total dataset representing customers who made previous purchases, oversampling was found to yield the best predictions. As KNN demonstrated superior results across all metrics, especially for AUC and precision in this study, it was selected as the model. To conclude, using a KNN model could reduce customer resources for World Plus which would in return reduce operational costs. This research could be expanded by choosing a reduced set of features and implementing alternative techniques for handling missing values and encoding as well as exploring different models.

REFERENCE

[1] Ahmad, A.K., Jafar, A. & Aljoumaa, K. 2019, "Customer churn prediction in telecom using machine learning in big data platform", Journal of big data, vol. 6, no. 1, pp. 1-24.

[2] Boustani, N., Emrouznejad, A., Gholami, R., Despic, O. & Ioannou, A. 2023, "Improving the predictive accuracy of the cross-selling of consumer loans using deep learning networks", Annals of operations research.

[3] Buuren, S.v. & Groothuis-Oudshoorn, K. 2011, "mice: Multivariate Imputation by Chained Equations in R", Journal of Statistical Software, vol. 45, no. 3.

[4] Donders, A.R.T., van der Heijden, Geert J.M.G., Stijnen, T. & Moons, K.G.M. 2006, "Review: A gentle introduction to imputation of missing values", Journal of clinical epidemiology, vol. 59, no. 10, pp. 1087-1091.

[5] Kwiatkowska, J. 2019, "CROSS-SELLING AND UP-SELLING IN A BANK", Copernican Journal of Finance & Accounting, vol. 7, no. 4, pp. 59-70.

[6] Lalwani, P., Mishra, M.K., Chadha, J.S. & Sethi, P. 2022, "Customer churn prediction system: a machine learning approach", Computing, vol. 104, no. 2, pp. 271-294.

[7] Moeyersoms, J. & Martens, D. 2015, "Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector", Decision Support Systems, vol. 72, pp. 72-81.

[8] Raschka, S. & Mirajalili, V. 2017, Python machine learning: machine learning and deep learning with Python, sci-kit-learn, and TensorFlow, Second edn, Packt, Birmingham.

[9] Sekeroglu, A. 2021, Impacts of Feature Selection Techniques in Machine Learning Algorithms for Cross Selling: A Comprehensive Study for Insurance Industry, doi:https://doi.org/10.13140/RG.2.2.24254.41284.

[10] Shearer, C. 2000, “The CRISP-DM model: The new blueprint for data mining”, Journal of Data Warehousing, 5(4), 13–22.

[11] Geng, Z., Hu, X., Zhu, Q., Han, Y., Xu, Y. & He, Y. 2018, "Pattern recognition for water flooded layer based on ensemble classifier", IEEE, pp. 164.

APPENDIX

Table 4: Evaluation of the confusion matrix for all models

No.	Model Name	Balancing Method	Accuracy	Precision	Recall	F1	AUC
1	Desicion Tree	both(p=0,5)	0,781	0,388	0,832	0,529	0,846
2	Desicion Tree	over(p=0,5)	0,775	0,382	0,838	0,525	0,846
3	Desicion Tree	smote(2,2)	0,781	0,388	0,832	0,529	0,852
4	Desicion Tree	smote(2,5,1,5)	0,779	0,386	0,836	0,528	0,846
5	Desicion Tree	under(p=0,5)	0,779	0,386	0,836	0,528	0,846
6	KNN	balance	0,974	0,926	0,894	0,910	0,941
7	KNN	over(p=0,5)	0,976	0,934	0,903	0,918	0,946
8	KNN	smote(2,2)	0,962	0,901	0,830	0,864	0,902
9	KNN	smote(2,5,1,5)	0,961	0,957	0,771	0,854	0,882
10	KNN	under(p=0,5)	0,969	0,911	0,878	0,894	0,932
11	LogRes	(tuning)	0,882	0,598	0,618	0,608	0,880
12	LogRes	both(p=0,5)	0,820	0,440	0,785	0,564	0,885
13	LogRes	over(p=0,5)	0,820	0,439	0,786	0,563	0,885
14	LogRes	smote(2,2)	0,846	0,487	0,741	0,588	0,884
15	LogRes	smote(2,5,1,5)	0,823	0,444	0,782	0,566	0,884
16	LogRes	under(p=0,5)	0,820	0,439	0,787	0,564	0,885
17	Random Forest	both(p=0,5)	0,851	0,498	0,761	0,602	0,893
18	Random Forest	over(p=0,5)	0,852	0,500	0,761	0,603	0,893
19	Random Forest	smote(2,2)	0,873	0,557	0,711	0,624	0,892
20	Random Forest	smote(2,5,1,5)	0,856	0,509	0,754	0,608	0,892
21	Random Forest	under(p=0,5)	0,852	0,500	0,761	0,603	0,893
22	RF after tuning(ntree=100)	both(p=0,5)	0,876	0,569	0,681	0,620	0,887
23	RF after tuning(ntree=100)	over(p=0,5)	0,902	0,717	0,562	0,630	0,880
24	RF after tuning(ntree=100)	under(p=0,5)	0,832	0,461	0,786	0,581	0,889
25	RF after tuning(ntree=100)	smote(2,5,1,5)	0,874	0,564	0,668	0,612	0,882
26	RF after tuning(ntree=100)	smote(2,2)	0,885	0,606	0,636	0,620	0,885
27	SVM	both(p=0,5)	0,803	0,417	0,824	0,553	0,881
28	SVM	over(p=0,5)	0,800	0,413	0,829	0,551	0,881
29	SVM	smote(2,2)	0,853	0,503	0,749	0,602	0,872
30	SVM	smote(2,5,1,5)	0,811	0,427	0,813	0,560	0,877
31	SVM	under(p=0,5)	0,799	0,410	0,825	0,548	0,880
32	xgboost	balance	0,839	0,474	0,794	0,594	0,881
33	xgboost	over(p=0,5)	0,871	0,550	0,681	0,609	0,880
34	xgboost	smote(2,2)	0,886	0,608	0,642	0,625	0,886
35	xgboost	smote(2,5,1,5)	0,895	0,662	0,595	0,627	0,885
36	xgboost	under(p=0,5)	0,818	0,436	0,790	0,562	0,882