Masters Programmes: Assignment Cover Sheet

Student Number:	5586298, 5503558, 5504970, 5545112, 2028065, 5575567
Module Code:	IB9BW0
Module Title:	Analytics in Practice
Submission Deadline:	12:00 (UK time) Wednesday 6 December 2023
Date Submitted:	Tuesday 5 December 2023
Word Count:	1996
Number of Pages:	14
Question Attempted: *(question number/title, or description of assignment)*	
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

1. Introduction

This report addresses the need of World Plus private bank, aiming to identify the target customers anticipated to make purchases through various communication platforms. Our analytics consultancy firm focuses on presenting a comprehensive data mining CRISP-DM approach tailored to company's requirements. Leveraging the provided dataset, we meticulously executed all six stages, encompassing data cleaning and preparation, alongside deploying diverse predictive models including Logistic Regression, Decision Trees, Super Vector Machine (SVM), and Random Forest. Our model evaluation has yielded valuable insights for implementation by World Plus.

2. Literature Review

2.1. "Exploring data sampling techniques for imbalanced classification problems" from SPIE

This paper examines the impact of imbalanced data in classification tasks, comparing ten sampling methods on real-world datasets. It uses metrics like precision, kappa coefficient, and G-measure to evaluate these methods. The study suggests balancing training data equally between majority and minority classes and training models on 60% of this data. It concludes by recommending SMOTE as the best methods for improving model performance, with random over and under sampling as viable alternatives. This research assists in selecting effective data balancing techniques for better model accuracy.

2.2. "Predicting customer demand for remanufactured products: A data-mining approach" from European Journal of Operational Research

This paper introduces a data mining prediction approach aimed at establishing a robust demand prediction model for remanufactured products. Following the CRISP-DM framework, the study employs three regression tree models—CART, M5, and RF. The paper provides guidance in the data preparation stage, offering insights into handling missing values with a set threshold of 5%. Additionally, a 60-40% split for training and test data was implemented. The importance of tuning is underscored, comparing machine learning models to determine optimal parameters that showed the superior performance of RF.

2.3. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead" from Nature Machine Intelligence

This paper suggests caution with black box models that can be explained after the fact and emphasizes the importance of inherently interpretable models for decision-making, aligning

with recent research trends. Managers may prioritize models aligned with business goals and understandable content, even if it sacrifices some predictive accuracy. Also, there is often no significant difference in performance between complex classifiers and simpler classifiers after preprocessing with structured data and meaningful features. Therefore, we will not use all significant features in information gain but select more meaningful features.

2.4. "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology"

The paper describes an implementation of data mining project founded upon CRISP-DM methodology aimed at forecasting the likelihood of client subscribing to new product offered by bank. The study incorporated three distinct data mining techniques including Naïve Bayes (NB), Decision Trees (DT) and Support Vector Machines (SVM). The empirical findings indicated that, among the three techniques, SVM outperformed its counterparts in predicting accuracy. This was evaluated based on the result of ROC and lift analysis. Given the similarities, we can adopt a similar machine learning model and evaluation methods.

2.5. "Surveying Techniques for High-Class Imbalance in Big Data"

This paper was selected for its comprehensive review of methods addressing class imbalances, a challenge that directly pertains to World Plus's need to predict customer conversion with unbalanced data. It will support our project's data preparation phase, where we aim to correct imbalances in the dataset to improve the accuracy of our predictive model. The paper's findings endorse the use of data sampling techniques, providing evidence for the effectiveness of combining Random Over-Sampling with varying split proportions, which justifies our approach of using 50% and 85% splits to mitigate prediction bias.

2.6. "Review of Methods for Handling Class-Imbalanced in Classification Problems"

The paper assesses some cutting-edge methods for handling class-imbalance classification problems. A variety of methods are used on the imbalanced datasets, such as deep learning, context-sensitive learning, algorithm-level methods, and data-level methods with oversampling, undersampling, and hybrids are used for the training datasets. Class-imbalance issues are handled with techniques like deep learning and cost-sensitive learning and various evaluation metrics to evaluate the models. Hybrid algorithms are more effective than resampling methods, but they are computationally more expensive and difficult to implement.

3. CRISP-DM Methodology

3.1. Understanding the Business Problem

Word Plus is a private bank company offering a range of financial products and services including loans, investment options, savings accounts, and credit products. The bank leverages various communication platforms, call centres, live chat, email, and social media to effectively promote their products to existing customers. The management team is trying to strategically utilize these communication channels to the target customers, aiming to enhance both sales and cost savings. In pursuit of this strategy, our consultancy team employed a CRISP-DM framework to identify potential customers for bank's new term deposit product, ensuring a cost-effective approach aligned with business objectives.

3.2. Data Understanding and Preparation

For World Plus project, we developed a lead prediction system with a dataset of 220,000 records and 16 variables, encompassing customer demographics and financial details. A comprehensive variable description is included in Appendix A. The initial phase of data preparation involved identifying and addressing quality issues. Notably, the 'Dependent' variable, expected to be binary, had -1 errors, constituting 0.0536% of the data. Therefore, records were removed to enhance accuracy (Nguyen et al., 2020).
The 'Credit_Product' variable had 8.30% missing values, which were addressed through imputation using the 'mice' package 'Active' and 'Gender' variables were binary-transformed for consistency. Upon plotting the data, no extreme values or patterns were observed. Subsequently, all variables initially stored as integers or characters were converted to factors, as detailed in Appendix B.

The next significant step involved transforming categorical variables to fit the predictive model. During this process, a deliberation arose regarding the categorization of the "Account_Type" variable as nominal or ordinal. Analysis of the total number of customers for each type and their average account balance revealed no significant differences. Consequently, we designated it as a nominal variable. Following this decision, one-hot encoding was applied to all nominal variables, including "Occupation," "Channel_Code," and "Account_Type". Omitting "Customer ID" and "Region_code" variables, deemed redundant and unlikely to impact our analysis, concluded the preprocessing steps.

3.3. Data Partitioning and Balancing

In the data partitioning and balancing phase, we initiated setting a random seed for reproducibility, followed by splitting the dataset into a 60% training set and a 40% testing set (Nguyen et al., 2020). This approach diverges from a traditional 50-50 split, aligning with the methodology of middle value (Leevy et al., 2018). However, upon assessing the proportion of the target variable in the training set, a noteworthy observation emerged - a significant majority, 85.2% of customers, did not purchase the product, while only 14.2% completed a purchase. To address this imbalance, we applied various sampling techniques, including oversampling, undersampling, a combination of both and SMOTE, to achieve dataset balance. However, SMOTE balancing model did not sufficiently improve efficiency (Sui et. al, 2019). Therefore, bothsampling method was deemed most effective, adjusting data by adding and removing records to maintain diversity and quality while achieving a balanced class distribution (Rawat and Mishra, 2022).

To select the most informative attributes for modelling stage, we computed information gain for each attribute, enabling subset creation (Rudin, 2019). This strategic feature selection significantly bolstered model accuracy by reducing dimensionality, thus enhancing both mining efficiency and result comprehensibility (Ganganwar, 2012). The result highlighted "Registration", "Age" and “Channel_Code_X1” as the three most crucial features. Variables with zero information gain were eliminated, as detailed in Appendix C.

3.4. Model Building

3.4.1 Model Methods and Results

We employed four supervised machine learning models—Support Vector Machine (SVM), Decision Tree, Random Forest (RF), and Logistic Regression—to predict targeted customers, as these are commonly used classification algorithms for handling datasets. Notably, SVM exhibited a longer runtime, necessitating sampling for code execution. Figure 2 vividly presents the matrices corresponding to each model.

Model Building

Model Name	**Training Ratio**	**Feature Filter**	**Sampling Technique**	**Model Technique**	**Accuracy**	**Precision**	**Recall**	**F1**
SVM	0.6	removed info gain values > 0.001	both (p = 0.35)	Radial	0.8963	0.66406	0.6024	0.6317
Random Forest	0.6	Used all variables	both (p = 0.35)	ntree = 900	0.893	0.64253	0.6203	0.6312
Decision Tree	0.6	removed info gain values > 0.001	both (p = 0.35)	NA	0.7998	0.4041	0.7498	0.5252
Log Reg	0.7	Note: Threshold = 0.5	both (p = 0.50)	NA	0.8263	0.4481	0.7512	0.5614

Figure 1 Result Comparison of All Models

4. Model Evaluation

In the model evaluation stage, diverse metrics were employed to assess the accuracy and reliability of predictions on unseen test data. One of the most common measures is accuracy, which is equal to 1- Error Rate. While accuracy, representing the percentage of positive classifications, is commonly used, its reliability diminishes in imbalanced datasets (Bekkar et al., 2013). Given the dataset's imbalance, where positive instances are fewer, accuracy becomes a misleading measure. In World Plus case, accurately identifying likely buyers is cost-effective by reducing resources spent on disinterested customers, making precision a crucial factor.

Considering the potential cost of targeting non-interested customers (False Positives), a strategy favouring high precision is advocated. In this scenario, a model that precisely identifies a smaller yet more confident number of highly probable leads is preferred over one capturing more leads but potentially including more false positives (High Recall).

Consequently, the F1 score emerged as the paramount measure in model evaluation. By striking a balance between recall and precision, the F1 score ensures the identification of as many positive instances as possible while mitigating the risk of false positives, aligning with World Plus's objectives and the challenges posed by the imbalanced dataset.

4.1. Model Improvement

During our model improvement phase, we explored hyperparameter tuning for every model we build, except for Logistic Regression. Several parameters were adjusted, such as the number of trees, node size, and variables for splitting were adjusted for Random Forest, and parameters like minimum bucket size and tree depth for Decision Tree and cost value for SVM. However, the tuned model did not show significant improvement over the original model in terms of predictive accuracy and computational efficiency. Based on the result of every model's matrix, we decided to proceed with the original Random Forest model, valuing its simplicity, balance result, and effectiveness. Appendix D presents the enhanced results of every model.

4.2. Model Comparison

Since there was no significant improvement after tuning, we decide to use the original Random Forest Model. Appendices D and E contained detailed tables, outlining measures values for each model both before and after the tuning process.

Furthermore, assessing model performance can be represented visually using a Receiver Operating Characteristics (ROC) graph, plotting False Positive Rate (FPR) against True Positive Rate (TPR) to show the trade-off between true and false positives (Moro et al., 2014). The Area Under Curve (AUC), quantifies the area beneath the ROC curve, with its values ranging from zero to one (Calders and Jaroszewicz, 2007). Random Forest exceeds all other models with a slightly higher AUC level of 08848.

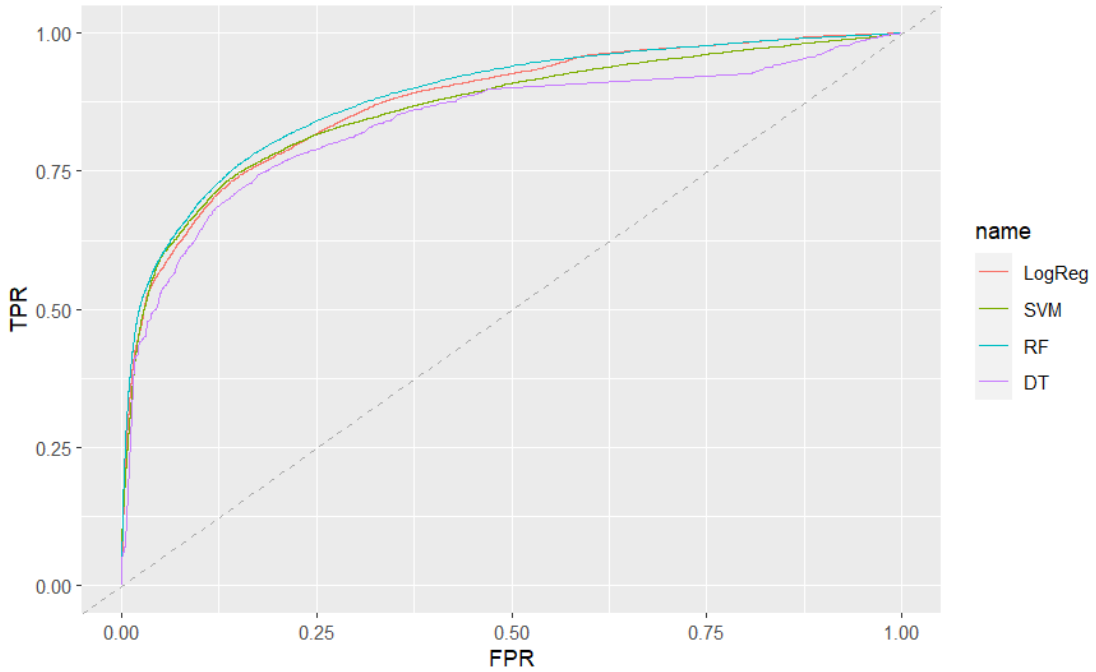


Figure 2: ROC graph

Model	AUC
Logistic Regression	0.8747
SVM	0.8637
Random Forest	0.8848
Decision Tree	0.8388

Figure 3: Table of Area Under the Curve (AUC)

4.3. Prediction Results

The cumulative gain chart assesses a model's ability to identify. In comparison to a random chance model, our model demonstrates superior performance. For instance, if World Plus aims to target the top 35% of customers, our model ensures that 80% of total customers who will make a purchase belong to this group.

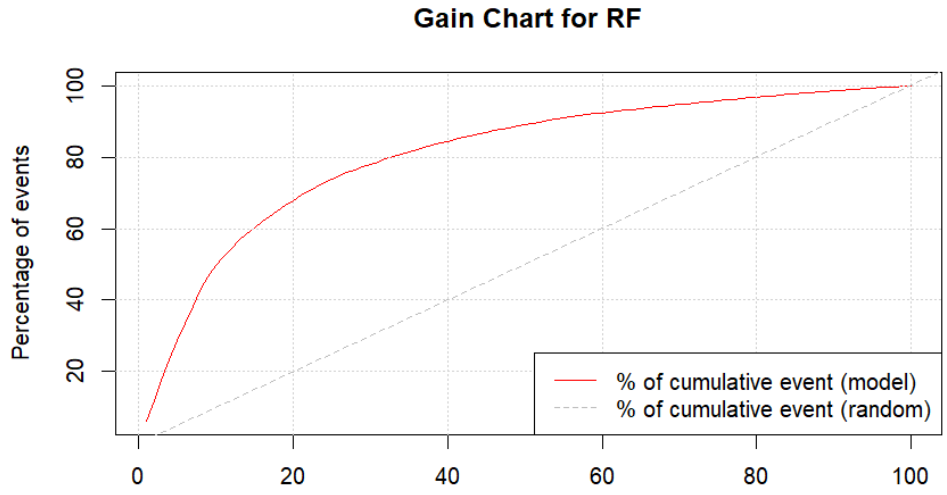


Figure4: Gain Chart

5. Conclusion

The data analysis project for World Plus private bank, using CRISP-DM, successfully developed a reliable lead prediction system with the Random Forest model outperforms other models in both accuracy and, more crucially, F1 score. This model is recommended for targeting potential customers for term deposit products, optimizing marketing resources.

In our pursuit of advancing World Plus's predictive analytics system, we plan to enhance World Plus's analytics by focusing on interpretable models, diverse data integration, and advanced feature analysis. The strategy includes seamless system integration, practical deployment, and a maintenance plan with regular updates and performance monitoring, aiming to support ongoing business growth and customer satisfaction.

6. References

Ayele, W.Y. (2020). Adapting CRISP-DM for Idea Mining. *International Journal of Advanced Computer Science and Applications*. Available at https://thesai.org/Publications/ViewPaper?Volume=11&Issue=6&Code=IJACSA&SerialNo=3 (Accessed 5 December 2023)

Bekkar, M., Kheliouane Djemaa, H., Taklit, A. and Alitouche (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. [online] 3(10). Available at: https://core.ac.uk/download/pdf/234677037.pdf (Accessed 4 Dec. 2023)

Broby, D. (2022). The Use of Predictive Analytics in Finance. *SSRN Electronic Journal*. Available at https://www.sciencedirect.com/science/article/pii/S2405918822000071 (Accessed 5 December 2023)

Calders, T. and Jaroszewicz, S. (2007). Efficient AUC Optimization for Classification. *Lecture Notes in Computer Science*, 4702, pp.42–53.

Choirunnisa, S., Lianto, J. (2022). Hybrid Method of Undersampling and Oversampling for Handling Imbalanced Data. IEEE Conference Publication. Available from: https://ieeexplore.ieee.org/abstract/document/8864335 (Accessed 20th November 2023)

Davis, Jesse and Goadrich, Mark. The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning. Available at: https://www.researchgate.net/publication/215721831_The_Relationship_Between_Precision-_Recall_and_ROC_Curves (Accessed 30 November 2023)

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4). Available at https://www.researchgate.net/publication/292018027_An_overview_of_classification_algorithms_for_imbalanced_datasets (Accessed 5 December 2023)

Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A. and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*. Available at https://www.researchgate.net/publication/328678792_A_survey_on_addressing_high-class_imbalance_in_big_data (Accessed 5 December 2023)

Moro, S., Laureano, R. and Cortez, P. (2014). Using data mining for bank direct marketing: an application of the CRISP-DM methodology. Uminho.pt. [online] doi: https://doi.org/978-90-77381-66-3

RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.

Singh Rawat, S. and Kumar Mishra, A. (2022). Review of Methods for Handling Class-Imbalanced in Classification Problems. [online] Available from: https://arxiv.org/pdf/2211.05456.pdf (Accessed 4 Dec. 2023)

Sui, Y., Zhang, X., Huan, J., Hong, H. (2019) Exploring data sampling techniques for imbalanced classification problems. Proc. SPIE 11198, Fourth International Workshop on Pattern Recognition, 1119813. Available from: https://0-doi-org.pugwash.lib.warwick.ac.uk/10.1117/12.2540457 (Accessed 20th November 2023)

Van Nguyen, T., Zhou, L., Chong, A.Y.L., Li, B. and Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research*, 281(3), pp.543–558.

7. Appendix

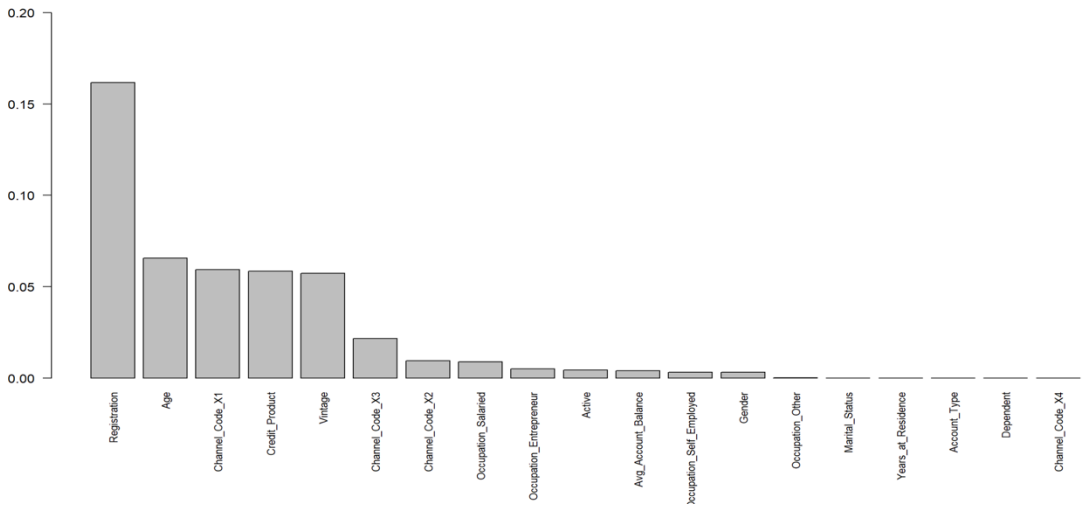
Appendix A: *Data Dictionary*

Variable	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	the duration in the current residence (in years)
Occupation	occupation type of the customer
Channel_Code	acquisition channel code used to reach the customer when they opened their bank account
Vintage	the number of months that the customer has been associated with the company.
Credit_Product	if the customer has any active credit product (home loan, personal loan, credit card etc.)
Avg_Account_Balance	average account balance for the customer in last 12 months
Account_Type	account type of the customer with categories Silver, Gold and Platinum
Active	if the customer is active in last 3 months
Registration	whether the customer has visited the bank for the offered product registration (1 = yes; 0 = no)
Target	whether the customer has purchased the product, 0: Customer did not purchase the product 1: Customer purchased the product

Appendix B: *Summary table of variable type*

Variable	Data types (Categorical/ Numerical)	Change to factor (Yes/No)
ID	Numerical integer	No
Gender	Categorical- character	Yes
Age	Numerical -integer	No
Dependent	Numerical -integer	Yes
Marital_Status	Numerical -integer	Yes
Region_Code	Categorical- character	No
Years_at_Residence	Numerical -integer	Yes
Occupation	Categorical- character	Yes
Channel_Code	Categorical- character	Yes
Vintage	Numerical -integer	No
Credit_Product	Categorical- character	Yes
Avg_Account_Balance	Numerical -integer	No
Account_Type	Categorical- character	Yes
Active	Categorical- character	Yes
Registration	Numerical -integer	Yes
Target	Numerical -integer	Yes

Appendix C: *Information gain plot for each attribute*



Appendix D: *Result Comparison of All Original Models*

Model Algorithms	Model Building					Stratified Sampled	Nodesize	Mtry
	Balancing (p = 0.35)	Accuracy	Precision	Recall	F1			
Decision Tree	Oversample	0.8561	0.51034	0.62831	0.56321	-	-	-
	Undersample	0.8865	0.61144	0.63416	0.62259	-	-	-
	Bothsample	0.8507	0.496	0.6799	0.5736	-	-	-
SVM (without information gain)	Oversample	0.8933	0.64964	0.60183	0.62482	0.25	-	-
	Undersample	0.8925	0.6455	0.60409	0.62411	0.5	-	-
	Bothsample	0.893	0.64842	0.60203	0.62437	0.25	-	-
SVM (with information gain)	Oversample	0.8961	0.66379	0.60049	0.63056	0.25	-	-
	Undersample	0.8931	0.6483	0.604	0.6254	0.5	-	-
	Bothsample	0.8951	0.65918	0.59977	0.62808	0.25	-	-
Random Forest	Oversample	0.8968	0.66636	0.60275	0.63296	-	-	-
	Undersample	0.8878	0.61513	0.64114	0.62786	-	-	-
	Bothsample	0.8930	0.64253	0.6203	0.63122	-	-	-
Logistic Regression	Oversample	0.8773	0.57839	0.62523	0.6009	-	-	-
	Undersample	0.879	0.58481	0.62215	0.6029	-	-	-
	Bothsample	0.8787	0.58383	0.62266	0.60262	-	-	-

Appendix E: *Result Comparison after model improvement through tuning*

Model Algorithms	Model Improvement (Tuned)					Stratified Sampled	Nodesize	Mtry
	Balancing (p = 0.35)	Accuracy	Precision	Recall	F1			
Decision Tree	Oversample	0.8949	0.66592	0.57801	0.61886	-	-	-
	Undersample	0.8948	0.66607	0.57719	0.61846	-	-	-
	Bothsample	0.8948	0.66607	0.57719	0.61846	-	-	-
SVM (without information gain)	Oversample	0.8923	0.644	0.6053	0.6241	0.05	-	-
	Undersample	0.8918	0.64141	0.60573	0.62306	0.2	-	-
	Bothsample	0.8926	0.64641	0.60162	0.62321	0.06	-	-
SVM (with information gain)	Oversample	0.8839	0.60107	0.63673	0.61838	0.06	-	-
	Undersample	0.8923	0.64505	0.60234	0.62296	0.15	-	-
	Bothsample	0.8839	0.60107	0.63673	0.61838	0.06	-	-
Random Forest	Oversample	0.9036	0.73693	0.53972	0.6231	-	1	7
	Undersample	0.876	0.5694	0.6575	0.6103	-	7	18
	Bothsample	0.893	0.64253	0.6203	0.63122	-	1	4
Logistic Regression	Oversample	-	-	-	-	-	-	-
	Undersample	-	-	-	-	-	-	-
	Bothsample	-	-	-	-	-	-	-