Phase II Report

Predicting Client Subscription to Term Deposit

Authors: Dikshya Niraula, Keith Sheridan, and Ting Zhang

Date: December 13, 2023

1. Introduction

Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephone marketing, and digital marketing.

Telephone marketing campaigns remain one of the most effective ways to reach people. However, they require significant investment as large call centers are hired to execute these campaigns. Thus, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted. The data for this project is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit.

Project Goals

1. Develop an accurate classification model to determine which clients are most likely to subscribe to the term deposit plans.
2. Increase the successful subscription rate among the target clients to the term deposit plan.

2. Methodology

Data

The data is related to the direct marketing campaigns of a Portuguese banking institution, conducted from May 2008 to June 2013. The marketing campaigns were based on phone calls. Contacts can be categorized as either inbound or outgoing based on the party (client or contact center) who initiated the interaction. The dataset contains 45,211 observations with 16 input variables, 10 categorical and 6 numeric. The table below gives a description of each variable in the dataset.

Table 1. Summary of each feature of the dataset.

Variable Name	Type	Description
age	Predictor - Numeric	Age of the client
job	Predictor - Categorical	Type of job

marital	Predictor - Categorical	Marital status
education	Predictor - Ordinal	Education level

Variable Name	Type	Description
default	Predictor - Binary	Has credit in default?
balance	Predictor - Numeric	Average yearly balance
housing	Predictor - Binary	Has a housing loan?
loan	Predictor - Binary	Has a personal loan?
campaign	Predictor - Numeric	Number of contacts performed during this campaign and for this client
pdays	Predictor - Numeric	Number of days that passed after the client was last contacted from a previous campaign
previous	Predictor - Numeric	Number of contacts performed before this campaign
poutcome	Predictor - Categorical	Outcome of the previous marketing campaign
Is_subscription	Response - Binary	Has the client subscribed to a term deposit?

Note: The * variables are related to the last contact of the current campaign.

The target variable of our model is "is_subscription", which indicates whether a client subscribed to a term deposit. Techniques to address the class imbalance issue will be discussed later in the paper.

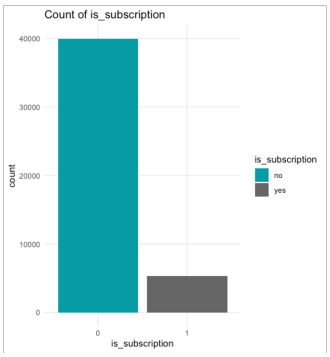


Figure 1. Distribution of target variable.

Variable Imputation

During our analysis, it was determined that the variables job, education, contact, and poutcome contained missing values. We decided to impute all missing values with the corresponding mode of the categorical predictor.

Table 2. Summary of imputed value of the variables containing missing values.

Variable	Proportion Missing	Imputed Value
job	0.64%	Mode: blue-collar
education	4.11%	Mode: secondary
contact	28.80%	Mode: cellular
poutcome	0.01%	Mode: failure

Variable Transformation

Yeo-Johnson transformations were applied to the numeric variables: age, balance, duration, campaign, and pdays. They helped improve the skewness and kurtosis of the distributions.

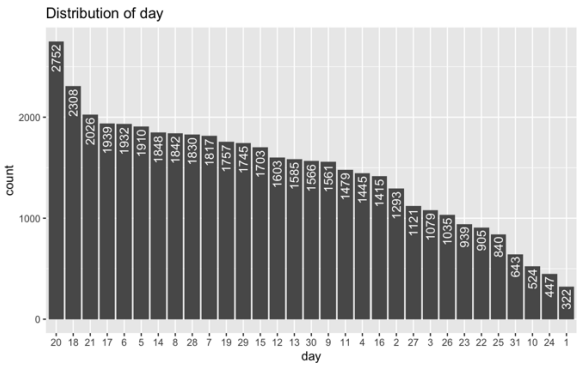
Initial Variable Importance Measures

Variable importance was measured by computing appropriate t- and chi-square statistics as well as p-values.

Table 3. Summary of the initial variable importance.

Variable Importance		
Variable	**p-value**	**ROC Area (num. only)**
duration	0	0.807
campaign	$3.728 \cdot 10^{-112}$	0.572
pdays		0.593
previous	$1.355 \cdot 10^{-71}$	0.602
balance	$4.383 \cdot 10^{-23}$	0.590
age	$1.597 \cdot 10^{-5}$	0.507
month	0	
poutcome	0	
contact	$1.251 \cdot 10^{-225}$	
housing	$2.918 \cdot 10^{-192}$	
job	$3.337 \cdot 10^{-172}$	
day	$6.896 \cdot 10^{-102}$	
education	$1.626 \cdot 10^{-51}$	
loan	$1.665 \cdot 10^{-47}$	
marital	$2.145 \cdot 10^{-43}$	
default	$2.453 \cdot 10^{-6}$	

day Variable Consolidation



The day variable (Figure 2) was consolidated using a Profit-Driven Decision Tree model. The modeling process resulted in the day variable being partitioned into five separate groups (Figure 3). Consolidation had an overall positive impact on the performance, notably leading to the significance of day variable in the model.

Figure 2. Distribution of the day variable.

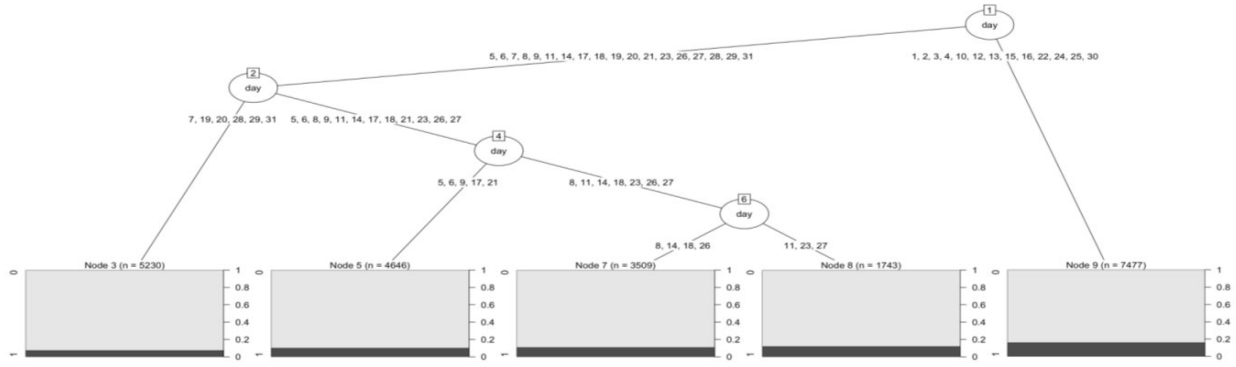


Figure 3. Profit-driven decision tree model of consolidation for day variable.

3. Model Comparison with Class Imbalance

The following models were implemented and assessed on the dataset with class imbalance: Decision Tree, Logistic Regression, Artificial Neural Network (ANN), and Random Forest. An alternative cut off was used to handle class imbalance (except Random Forest) and each model was evaluated based on its ability to maximize the F-score.

Decision Tree

The optimal Decision Tree resulted in 19 leaves. It employed a profit driven approach with cp threshold 0.001 and utilized 10 input variables: duration, month, poutcome, housing, age, pdays, job, previous, day, and balance. A maximum F-score of 0.53612820 was achieved.

Logistic Regression

The Logistic Regression, with consistent variable selection methods, selected an additional input variable contact.NA. A maximum F-score of 0.5334116 was achieved but did not surpass the decision tree.

Artificial Neural Network

The ANN comprised a hidden layer with 6 neurons and a hyperbolic tangent activation function. A maximum F-score of 0.577933 was achieved when incorporating inputs selected from logistic regression.

Random Forest

The Random Forest model, employing internal down-sampling, outperformed other models, achieving a maximum F-score of 0.5801775. The two most important variables, identified through both accuracy and Gini index metrics, were duration and month.

Table 4. Summary of the performance of the four models.

Model Comparison with Class Imbalance		
Model	**F-score**	**Area Under Curve (AUC)**
Decision Tree	0.5361	0.8752
Logistic Regression	0.5334	0.9130
Artificial Neural Network	0.5779	0.9224
Random Forest	0.5802	0.9310

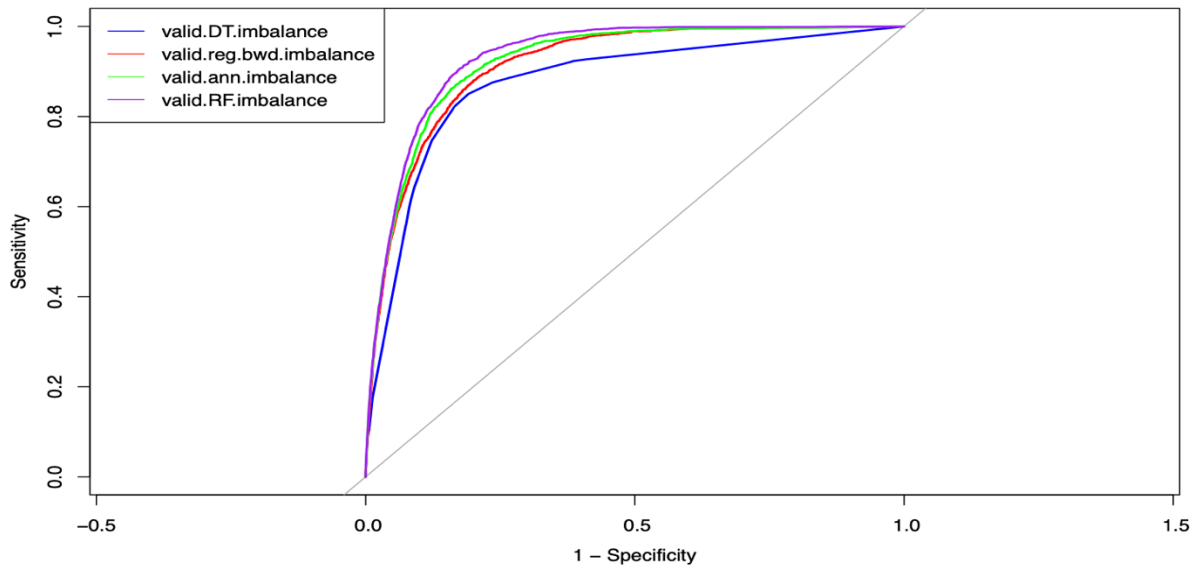


Figure 4. Comparison of ROC Curves.

4. Handling Class Imbalance

Four balancing techniques - undersampling, oversampling, both-way sampling, and rose sampling were evaluated. Each technique was systematically applied to the four models: Decision Tree, Logistic Regression, Artificial Neural Network (ANN), and Random Forest.

Table 5. Summary of each model performance based on decision tree modeling.

Decision Tree Model Comparison		
Model	**F-score**	**Area Under Curve (AUC)**
Decision Tree.Imbalance	0.5361	0.8752
Decision Tree.Over	0.5377	0.8838
Decision Tree.Under	0.5281	0.8867
Decision Tree.Both	0.5241	0.8759
Decision Tree.Rose	0.5325	0.8968

Table 6. Summary of each model performance based on logistic regression modeling.

Logistic Regression Model Comparison		
Model	**F-score**	**Area Under Curve (AUC)**
Logistic Regression.Imbalance	0.5334	0.9130
Logistic Regression.Over	0.5406	0.9140
Logistic Regression.Under	0.5366	0.9109
Logistic Regression.Both	0.5323	0.9126
Logistic Regression.Rose	0.5306	0.9129

Table 7. Summary of each model performance based on ANN modeling.

ANN Model Comparison		
Model	**F-score**	**Area Under Curve (AUC)**
ANN.Imbalance	0.5779	0.9224
ANN.Over	0.5667	0.9222
ANN.Under	0.5209	0.9105
ANN.Both	0.5474	0.9151

ANN.Rose	0.5357	0.9134

Table 8. Summary of each model performance based on random forest modeling.

Random Forest Model Comparison		
Model	**F-score**	**Area Under Curve (AUC)**
Random Forest.Imbalance	0.5802	0.9312
Random Forest.Over	0.6041	0.9293
Random Forest.Under	0.5442	0.9271
Random Forest.Both	0.6055	0.9284
Random Forest.Rose	0.5921	0.9264

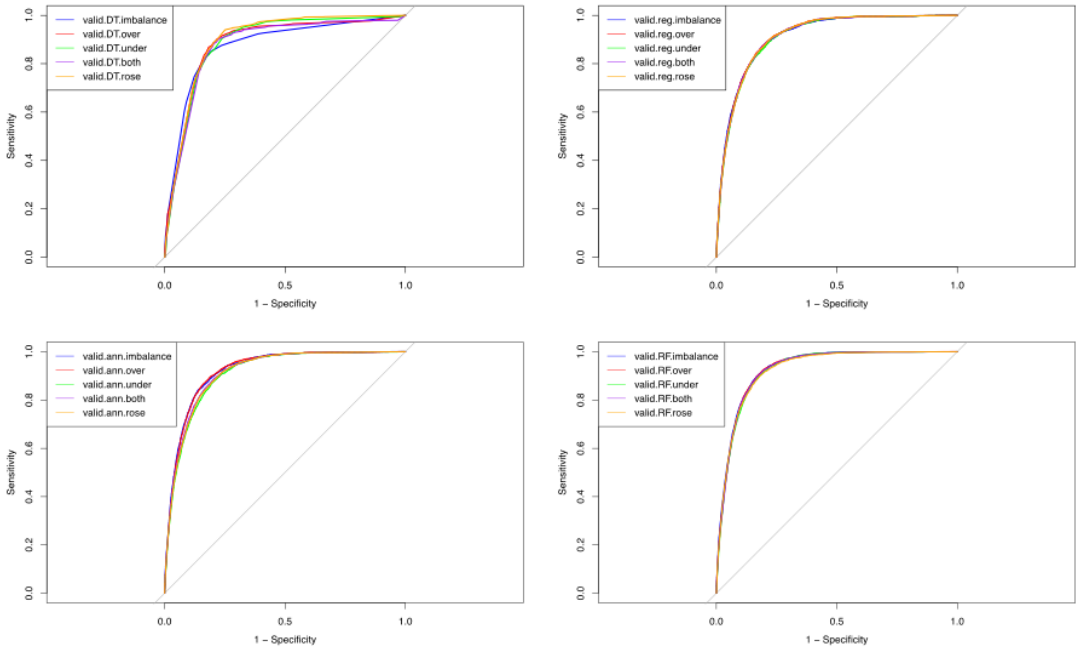


Figure 5. Comparison of ROC Curves.

Oversampling improved the F-score for the Decision Tree model, while both oversampling and undersampling led to increased F-score in the Logistic Regression model. The ANN model performed better without any balancing technique.

The top-ranking models based on both F-score and AUC comprised exclusively of Random Forest models. The consistent dominance of Random Forest models suggests that they are particularly well-suited for our dataset.

Table 9. Summary of top 5 models' performance based on F-Score.

Top 5 Models: F-Score	
Model	**Area Under Curve (AUC)**
Random Forest.Both	0.6055
Random Forest.Over	0.6041
Random Forest.Rose	0.5921
Random Forest.Imbalance	0.5802
ANN.Imbalance	0.5779

Table 10. Summary of top 5 models' performance based on AUC.

Top 5 Models: AUC	
Model	**Area Under Curve (AUC)**
Random Forest.Imbalance	0.9312
Random Forest.Over	0.9293
Random Forest.Both	0.9284
Random Forest.Rose	0.9264
ANN.Imbalance	0.9227

5. Best Performing Model

The Random Forest-balanced with oversampling model, which performed the best on the train data, performed extremely well on the test data. The main aim here is to maximize the bank's profit by identifying as many people as possible that will use the bank's term deposit. The model successfully identified 89.92% of the people that chose to use this deposit.

From the feature importance, we can see that the most important features are duration and month of previous contact. Duration is the last contact duration. These suggest the previous contact with a customer is predictive of marketing success. This makes sense because this means the customer

has a good relationship with the bank, which makes them more likely to subscribe to the bank term deposit.

6. Future Considerations

Although Random Forest-balanced shows the best performance among the models, its precision value is still lower than expected. In the future, several enhancements can be made to improve the model performance, such as optimizing the model hyperparameters. We shall explore complex relationships within the data and incorporate interaction terms, where deemed appropriate. We shall investigate techniques that smoothly integrate with SMOTE to accommodate categorical data.