TIERS Information Technology Journal
Vol.4, No.1, June 2023, pp. 43~55
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v4i1.4355

Implementation of LightGBM and Random Forest in Potential Customer Classification

Laura Sari, Annisa Romadloni[2], Rostika Lityaningrum[3], Hety Dwi Hastuti[4]
laurasari@pnc.ac.id[1], annisa.romadloni@pnc.ac.id[2], nadhifa007@gmail.com[3],
[1-3]Informatic Engineering, Politeknik Negeri Cilacap, Central Java
[2]Accounting Financial Institutions, Politeknik Negeri Cilacap, Central Java

ABSTRACT

Classification is one of the data mining techniques that can be used to determine potential custumers. Previous research show that the boosting method, especially LGBM, produces the highest accuracy value of all models, namely 100%. Meanwhile, for the two bagging methods, Random Forest produced the highest accuracy compared to Extra Trees, namely 99.03%. The research uses the LGBM and Random Forest methods to classify potential customers. The results of this study indicate that in imbalance data the LightGBM method has better accuracy than the Random Forest, which is 85.49%, when the Random Forest is unable to produce a model. The SMOTE method used in this study affects the accuracy of the random forest but does not affect the accuracy of LightGBM. Over all the Accuracy, Recall, Specificity, and Precision values, Random Forest produces a good value compared to LightGBM on balanced data. Meanwhile, LightGBM is able to handle unbalanced data.

Keywords: LightGBM; Random Forest; Clasification;

Article Info

Accepted	:	14-06-2023
Revised	:	09-04-2023
Published Online	:	25-06-2023

Correspondence Author:

Laura Sari
Informatic Engineering,
Politeknik Negeri Cilacap,
Jl. Dr. Soetomo No.1, Karangcengis, Sidakaya, Kec. Cilacap Sel., Kabupaten Cilacap, Jawa Tengah 53212.
Email: laurasari@pnc.ac.id

1. INTRODUCTION

Identification of potential customers is intended so that promotions can be carried out efficiently and effectively. Classification is one of the data mining techniques that can be used to determine potential customers. Research [1] on the search for a model that can increase the efficiency of promotion by identifying the main characteristics of customers so as to increase success and help management to be better in utilizing available resources, as well as quality and affordable selection of potential customer data sets. This study compares three models, namely Naïve Bayes, Decision Tree, and SVM. After testing, the AUC value for NB was 0.870, DT was 0.868, and SVM was 0.938. Another study [2] classifies potential customers based on occupation, type of salary, tenor and age using the Naïve Bayes Classifier method. This study produces a high accuracy value of 97%. Research [3,4] uses K-Means and K-NN.

Ensemble learning is a method by which multiple algorithms are used together. The goal of ensemble learning is to improve accuracy rather than using just one algorithm. Siringoringo and Jaya in [5] applied the ensemble technique of SMOTE Bagging method applied to unbalanced data. Unbalanced data has a devastating impact on classification results where minority classes are often misclassified into majority classes. Conventional data mining algorithms are not equipped with the ability to work on unbalanced data, so the performance produced by conventional algorithms is often not optimal. The result of this study is that SMOTE Bagging performance has better performance than SVM, K-NN, Decision Tree algorithms with average AUC from SVM, K-NN, Decision Tree and SMOTE Bagging algorithms 0.638, 0.742, 0.770 and 0.895 respectively. Two bagging methods, Random Forest produced the highest accuracy compared to Extra Trees, namely 99.03%.

Similar research aims to find potential customers in successful banking telemarketing using meta-algorithms such as bagging for decision tree models and produce 98.7% accuracy results [3]. Another study comparing bagging and boosting techniques on the CART method to classify students' study periods. The study obtained the results that the CART algorithm classification model with boosting techniques has the best Accuracy value for unbalanced classes compared to CART and CART models with bagging techniques, which is 81.25% [5].

Research compared three ensemble methods namely boosting, bagging and stacking to classify diabetes. The research resulted that the boosting method can outperform bagging and stacking methods, especially Light Gradient Boosting which provides the highest accuracy of 99.25%. Light Gradient Boosting also provides 99% accuracy results [6]. LightGBM speeds up process time 20 times over the conventional Gradient Boosting Decision Tree training phase with the same accuracy. In the prediction of breast cancer patients using several machine learning techniques, namely XG Boost, Random Forest, and LightGBM. From this study, it was obtained that LightGBM is superior in term of accuracy and speed than other techniques [7].

This study used two ensemble techniques, namely boosting and bagging. In the boosting technique, LightGBM is used because of its good accuracy and speed [7,9]. While the bagging technique will be used Random Forest which method is the best in terms of accuracy compared to other bagging techniques [8-14]. In addition, a lot of data is available but a lot is out of balance. Though unbalanced data causes predictions to be biased. Therefore, this study aims to determine the performance of LightGBM and Random Forest methods on unbalanced data and balanced data. For this reason, the SMOTE method is also needed to balance data.

2. RESEARCH METHOD

This study used the Cross-Indutry Standard Process for Data Mining (CRISP-DM) method. CRISP-DM is a popular method for increasing the success of data mining projects. This methodology defines six sequential but flexible steps in building and deploying a data mining model for use on real problems and supporting business objectives. The CRISP-DM method allows iterating stages if needed. These stages are defining objectives (business understanding stage), then the data needs to be analyzed (data understanding stage) and processed (data preparation stage). The modeling stage is to build a model that fits the characteristics of the data. The resulting model can be used to predict the target value that represents the specified goal. Furthermore, the model is analyzed in the evaluation stage in terms of performance and usability. If the model obtained is not good enough, it can be remodeled. The best model can be implemented in real problems (deployment stage) [6]. The process is sawn in the Figure 1.

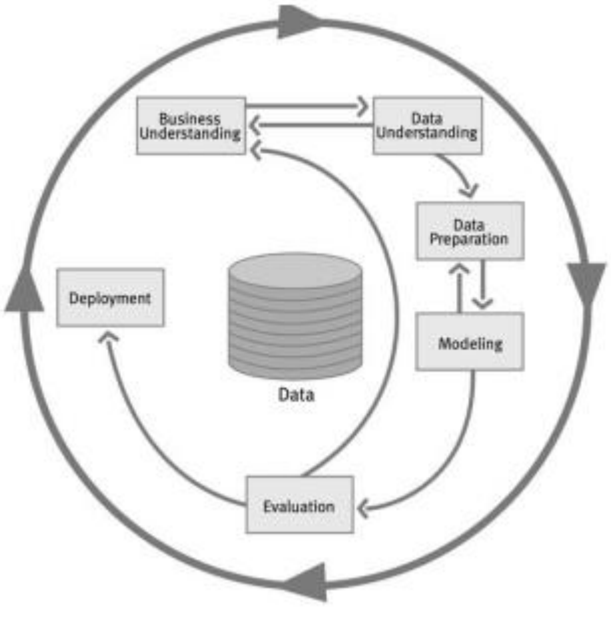

Figure 1 Research Stages

2.1. Business Understanding

The model in this study will be applied to Superstore Marketing Campaign data obtained on the Kaggle website. The data is a sample of customer data from a superstore. The superstore plans to hold year-end discounts. They made a new offer to gold membership customers giving 20% off every purchase of $499 which is $999. on another day. Promotions will be delivered to active customers by telephone. Management feels that the best way to reduce promotion costs is to build a predictive model that will classify the customers who will receive the promotion. The purpose of this study is to create a model to predict the probability that a customer will give a positive response and the factors that influence customer response. The data mining model that will be used is the ensemble method, namely the Light Gradient Boosting Machine (Light GBM) and the Random Forest. While the tools used are RStudio.

2.2. Data Understanding

The data used in the classification process is in the form of public data taken from the Kaggle website [17]. The data is data obtained in last year's promotion at a superstore. Consists of 2240 customers presented in rows and 22 criteria presented in columns. Information about each column is presented in Table 1 below.

Table 1 Table of Data Description

Column	Information
Id	Unique ID of each customer
Year_Birth	Age of the customer Complain
Complain	1 if the customer complained in the last 2 years, 0 if no complained
Dt_Customer	Date of customer's enrollment with the company
Education	Customer's level of education
Status_Marital	Customer's marital status
Kidhome	Number of small children in customer's household
Teenhome	Number of teenagers in customer's household
Income	Customer's yearly household income
MntFishProducts	The amount spent on fish products in the last 2 year
MntMeatProducts	The amount spent on meat products in the last 2 year
MntFruits	The amount spent on fruits products in the last 2 year
MntSweetProducts	The amount spent on sweet products in the last 2 year
MntWines	The amount spent on wine products in the last 2 year
MntGoldProds	The amount spent on gold products in the last 2 year
NumDealsPurchases	Number of purchases made with discount
NumCatalogPurchases	Number of purchases made using catalog (buying goods to be shipped through the mail)
NumStorePurchases	Number of purchases made directly in stores
NumWebPurchases	Number of purchases made through the company's website
NumWebVisitsMonth	Number of visit to company's website in the last month
Recency	Number of days since the last purchase
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

It can be seen that there are columns containing categorical data, namely Education, Marital Status, and Response. The rest contains numbers only. Meanwhile, Id and Dt_Customer are considered to have unique values for each row of data and represent each individual. At the data understanding stage, data quality checks are carried out to understand the characteristics of the data. Examination of data quality consists of data types, the presence of missing data, outliers, duplication of data, and examining data patterns. Algorithm designs are written as follows:g categorical data, namely Education, Marital Status, and Response. The rest contains numbers only. Meanwhile, Id and Dt_Customer are considered to have unique values for each row of data and represent each individual. At the data understanding stage, data quality checks are carried out to understand the characteristics of the data. Examination of data quality consists of data types, the presence of missing data, outliers, duplication of data, and examining data patterns. Algorithm designs are written as follows:

Program Data Understanding

```
library(readxl)
library(outliers)
library(tidyverse)
library(ggplot2)
library(lattice)
library(caret)
library(GGally)
library(cowplot)
library(lightgbm)
library(Matrix)
library(dplyr)

view(superstore_datanew)
head(superstore_datanew)
summary(superstore_datanew)
str(superstore_datanew)
superstore_datanew$Education = as.factor(superstore_datanew$Education)
superstore_datanew$Marital_Status = as.factor(superstore_datanew$Marital_Status)
superstore_datanew$Response = as.factor(superstore_datanew$Response)
superstore_datanew$Complain = as.factor(superstore_datanew$Complain)
summary(superstore_datanew)
ggcorr(superstore_datanew, hjust = 1, layout.exp = 2, label = T, label_size = 2.9)
plot_categoric_features <- function(x){
     ggplot(superstore_data, aes_string(x, "Response")) +
         geom_boxplot() +
         coord_flip() }
plot_grid(
     plot_categoric_features("Income"),
     plot_categoric_features("Kidhome"),
     plot_categoric_features("Teenhome"),
     plot_categoric_features("Recency"),
     plot_categoric_features("MntWines"),
     plot_categoric_features("MntFruits"),
     plot_categoric_features("MntMeatProducts"),
     plot_categoric_features("MntFishProducts"),
     plot_categoric_features("MntSweetProducts"),
     plot_categoric_features("MntGoldProds"),
     plot_categoric_features("NumDealsPurchases"),
     plot_categoric_features("NumWebPurchases"),
     plot_categoric_features("NumStorePurchases"),
     plot_categoric_features("NumWebVisitsMonth"))

nearZeroVar(superstore_datanew)
superstore_datanew %>%
     select(is.numeric) %>%
     outlier()
colSums(is.na(superstore_data))
```

In R, function summary() used to view information of the data such as minimal, maximal, mean, median, and amount of data in each category. The output of the above function is shown in Table 2.

Table 2 Summary Data

```
Id                  Age               Education      Marital_Status
 Length:2240         Min.   : 27.00   2n Cycle  : 203   Married :864
 Class :character    1st Qu.: 46.00   Basic     :  54   Together:580
 Mode  :character    Median : 53.00   Graduation:1127   Single  :480
                     Mean   : 54.19   Master    : 370   Divorced:232
                     3rd Qu.: 64.00   PhD       : 486   Widow   : 77
                     Max.   :130.00                     Alone   :  3
                                                        (Other) :  4
     Income          Kidhome          Teenhome        Dt_Customer
 Min.   :  1730   Min.   :0.0000   Min.   :0.0000   Length:2240
 1st Qu.: 35303   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
 Median : 51382   Median :0.0000   Median :0.0000   Mode  :character
 Mean   : 52247   Mean   :0.4442   Mean   :0.5062
 3rd Qu.: 68522   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :666666   Max.   :2.0000   Max.   :2.0000
 NA's   :24
    Recency        MntWines         MntFruits     MntMeatProducts
 Min.   : 0.00   Min.   :   0.00   Min.   :  0.0   Min.   :   0.0
 1st Qu.:24.00   1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0
 Median :49.00   Median : 173.50   Median :  8.0   Median :  67.0
 Mean   :49.11   Mean   : 303.94   Mean   : 26.3   Mean   : 166.9
 3rd Qu.:74.00   3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0
 Max.   :99.00   Max.   :1493.00   Max.   :199.0   Max.   :1725.0

 MntFishProducts  MntSweetProducts  MntGoldProds    NumDealsPurchases
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
 1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000
 Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000
 Mean   : 37.53   Mean   : 27.06   Mean   : 44.02   Mean   : 2.325
 3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000
 Max.   :259.00   Max.   :263.00   Max.   :362.00   Max.   :15.000

 NumWebPurchases   NumCatalogPurchases NumStorePurchases NumWebVisitsMonth Respons
e
 Min.   : 0.000   Min.   : 0.000      Min.   : 0.00     Min.   : 0.000    0:1906
 1st Qu.: 2.000   1st Qu.: 0.000      1st Qu.: 3.00     1st Qu.: 3.000    1: 334
 Median : 4.000   Median : 2.000      Median : 5.00     Median : 6.000
 Mean   : 4.085   Mean   : 2.662      Mean   : 5.79     Mean   : 5.317
 3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.00     3rd Qu.: 7.000
 Max.   :27.000   Max.   :28.000      Max.   :13.00     Max.   :20.000

 Complain
 0:2219
 1:  21
```

Furthermore, it is evident that the data quality is subpar due to an imbalanced distribution and the presence of outliers. The function "outlier()" can be utilized to identify all the outliers in the data, and it reveals the existence of outliers in each variable. Additionally, through the aforementioned test, it becomes apparent that there are missing values (NA) in the Income column, specifically 24 instances of missing data. Figure 2 shows that the response variables with classification 0 tend to be more than classification 1, which is around 85.089%.

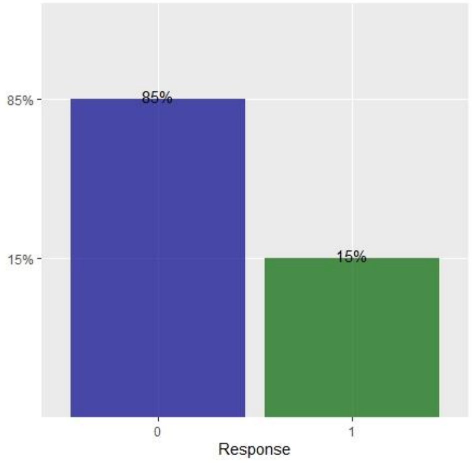

Figure 2. Proportion Response Data

2.3. Data Preparation

This stage includes selecting a subset of data, either in the form of columns or tables that are suitable for the purpose, improving data quality with data cleaning processes, and building data with data construct processes such as changing data types. Missing value handling is done by imputation using the median value. After this stage a data set will be generated that is ready to be modeled. Algorithm design are written as follows:

Program Data Preparation

```
superstore_data = superstore_datanew[, -nearZeroVar(superstore_datanew)]
superstore_data = superstore_datanew %>%
     filter(Age != 130,
            Income != 666666)
superstore_data$Income = ifelse(is.na(superstore_data$Income), ave(superstore_data$Income, FUN
= function(x) mean(x, na.rm = TRUE)), superstore_data$Income)
superstore_data   =   data.frame(superstore_data,   data.matrix(superstore_data[c("Education",
"Marital_Status")]) )
superstore_data = superstore_data[, -c(1,3,4,8)
view(superstore_data)
```

At this stage, the Education and Marital_Status variables are also recorded. The Education variable is divided into 5 categories, namely Basic = 1, 2n Cycle = 2, Graduation = 3, Master = 4, and PhD = 5. Meanwhile, the Marital_Status variable is divided into 8 categories, namely Absurd = 1, Alone = 2, Divorce = 3, Married = 4 , Single = 5, Together = 6, Widow = 7, YOLO = 8. Furthermore, outlier data that has checked in previous stage id deleted. The oulier removed is 130 in Age, 666666 in Income, 99 in Recency, 1493 in Mnt Wines, 199 in MntFruits, 1725 in MntMeatProducts, 259 in MntFishProducts, 263 in MntSweetProducts, 362 in MntGoldProds, 15 in NumDealsPurchases, 27 in NumWebPurchases, 13 in NumCatalogPurchases, and 20 in NumWebVisitMonth.

2.4. Modelling

In this study, the Light Gradient Boosting Machine (Light GBM) model is employed. Light GBM is a rapid and efficient gradient-boosting framework that relies on decision tree algorithms. It is utilized for various machine learning tasks such as ranking and classification. Gradient boosting, the underlying technique, is a machine-learning approach used for regression and classification problems. It generates an ensemble of weak prediction models, commonly decision trees, to create a predictive model. Light GBM has gained popularity in various machine-learning competitions and real-world applications due to its fasttraining speed, memory efficiency, and strong predictive performance. It can handle large datasets and high-dimensional feature spaces efficiently. The process is achieved by using a technique called Gradient-based One-Side Sampling (GOSS) that reduces the number of data instances used for training, focusing on the instances that have larger gradients. Light GBM grows trees in a leaf-wise manner, rather than the traditional level-wise approach. In the leaf-wise strategy, the algorithm grows the tree by adding nodes with the highest loss reduction, resulting in a more balanced and deeper tree structure. This approach can lead to better accuracy but may be prone to overfitting if not properly regularized. However, it is worth noting that the leaf-wise growth strategy and the potential for overfitting may require careful tuning of hyperparameters to obtain the best results for a given problem [4].

Furthermore, this study integrates the Random Forest technique, which is a commonly used machine learning method belonging to the ensemble learning category. Random Forest is applied in both classification and regression scenarios. It combines the predictions from multiple decision trees to produce more accurate and robust prediction [18]. Random Forest utilizes a technique called bootstrapping to create multiple subsets from the original dataset. Each subset is generated by randomly selecting data points from the original dataset, with replacements. This ensures that each subset has the same number of data points as the original dataset, although some points may be repeated or omitted. For each subset, a decision tree is constructed using a randomly chosen subset of features. At each node of the tree, a feature is selected to split the data based on a criterion like Gini impurity or information gain. This recursive process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples at a leaf node. Once all the decision trees are constructed, they make predictions independently. In classification tasks, the final prediction is determined by majority voting among the decision trees. In regression tasks, the final prediction is typically calculated as the average or median of the predictions from all the decision trees [7, 10]. In R programming, the LightGBM method uses the lightgbm package and Random Forest uses the Random Forest package. Both models are optimized using a model evaluation technique called K-Fold Validation. The data is divided into k parts and each part will be a data set in turn. So that every data has a chance into train and test data [16].

Unbalanced data is resolved by using the Synthetic Minority Oversampling Technique (SMOTE). The SMOTE technique is the most well-known technique for dealing with unbalanced data. This technique is similar to the oversampling technique, namely duplicating data from the minority class so that the sum is equal to the amount of data from the majority class. However, SMOTE does not only duplicate the same data, but SMOTE will create new samples that resemble the original data from the minority class so that the minority class becomes much more diverse [5].

2.5. Evaluation

Measuring the value of accuracy the model can be calculated using a formula:

$$\text{Accuracy}: \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$\text{Recall}: \frac{TP}{TP+FN} \times 100\%$$

$$\text{Precision}: \frac{TP}{TP+FP} \times 100\%$$

where *False Negative* (FN), *False Positive* (FP), *True Negative* (TN), dan *True Positive* (TP) [16]. Accuracy is the ratio of correct predictions for all data. Recall is the ratio of actual positive predictions to all positive actual data. Whereas precision is the ratio of true positive predictions compared to all positive actual data. Algorithm design are written as follows:

Program Modelling

```
#######Model LGBM######
set.seed(123)
smp_size = floor(0.8 * nrow(superstore_data))
train_ind = sample(seq_len(nrow(superstore_data)), size = smp_size)
train = superstore_data[train_ind, ]
val = superstore_data[-train_ind, ]
trainm = sparse.model.matrix(Response ~., data = train)
train_label = train[,"Response"]
valm = sparse.model.matrix(Response~., data= val)
val_label = val[,"Response"]
train_matrix = lgb.Dataset(data = as.matrix(trainm), label = train_label)
val_matrix = lgb.Dataset(data = as.matrix(valm), label = val_label)
valid = list(test = val_matrix)
params = list(max_bin = 10,
              learning_rate = 0.001,
              objective = "binary",
              metric = 'binary_logloss')
bst = lightgbm(params = params, train_matrix, valid, nrounds = 1000)
p = predict(bst, valm)
val$predicted = ifelse(p > 0.5,1,0)
confusionMatrix(factor(val$predicted), factor(val$Response))

######Smooting#######
data_train <- train %>%
            mutate(Response = as.factor(Response)) %>%
            mutate_if(is.character, as.factor)
data_train_smote <- SmoteClassif(form = Response ~ .,
                                  dat = data_train,
                                  C.perc = "balance",
                                  dist = "HVDM")
######Model Random Forest######
set.seed(123)
smp_size = floor(0.8 * nrow(smoote))
train_ind = sample(seq_len(nrow(smoote)), size = smp_size)
train = smoote[train_ind, ]
val = smoote[-train_ind, ]
set.seed(123)
smp_size = floor(0.8 * nrow(smoote))
train_ind = sample(seq_len(nrow(smoote)), size = smp_size)
train = smoote[train_ind, ]
test = smoote[-train_ind, ]
model_forest <- train(Response ~ ., data = train, method = "rf", trControl = ctrl)
saveRDS(model_forest, "model_rforest.RDS")
model_rforest <- readRDS("model_rforest.RDS")
model_rforest
model_rforest$finalModel
rf_table <- tibble(y = test$Response)
rf_prob <- predict(model_rforest, test, type = "prob")
rf_table$rf_prob_no <- round(rf_prob[,1],4)
rf_table$rf_prob_yes <- round(rf_prob[,2],4)
rf_table$rf_class <- factor(ifelse(rf_prob[,2] > 0.5, "1","0"))
rf_table$y=factor(rf_table$y)
cm_rf <- confusionMatrix(rf_table$rf_class, rf_table$y, positive = "1")
cm_rf
```

3. RESULTS AND DISCUSSION

The relationships between variables in the dataset are analyzed with the Pearson equation. The relationship between these variables can be seen in Figure 3. Relationships consist of direct relations and indirect relations. Direct relationships are in the range of 0 – 1 and indirect between 0 and -1. Some variables have a direct relationship above 0.5 namely Income, MntWines, MntFruits, MntMeatProduct, and NumCatalogPurchase. Indirect correlation below -0.5 i.e. variable: NumWebVisitsMonth.

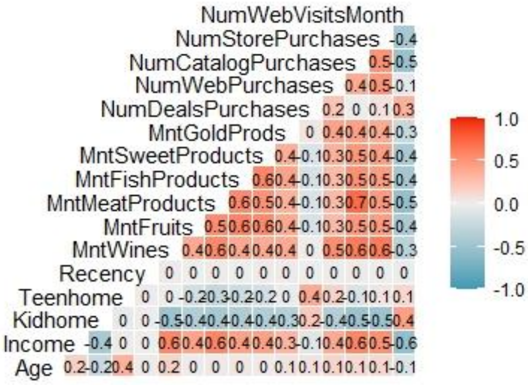

Figure 3 Matriks Corelation

Figure 4 show the relationship between each factor variable and the predictor. From the plot, we can deduce that individuals with higher income tend to exhibit a higher likelihood of giving a positive response. Additionally, those who enrolled in 2013 have the highest probability of giving a positive response. Moreover, individuals without more than one teenager or child are also inclined to give a positive response. Lastly, unmarried individuals and those who have completed their education are more likely to give a positive response. It was also observed that individuals who made more than five web purchases have a higher probability of giving a positive response. Similarly, those who made more than four catalog purchases (via mail) are more likely to give a positive response. Furthermore, individuals who have not made recent purchases are more likely to give a positive response. Additionally, individuals who spent more than $25 on Fruits, Sweet products, and Fish products are more inclined to give a positive response. Similarly, those who spent more than $50 on gold products are more likely to give a positive response. Moreover, individuals who spent more than $200 on meat products and more than $400 on wines are more likely to give a positive response.

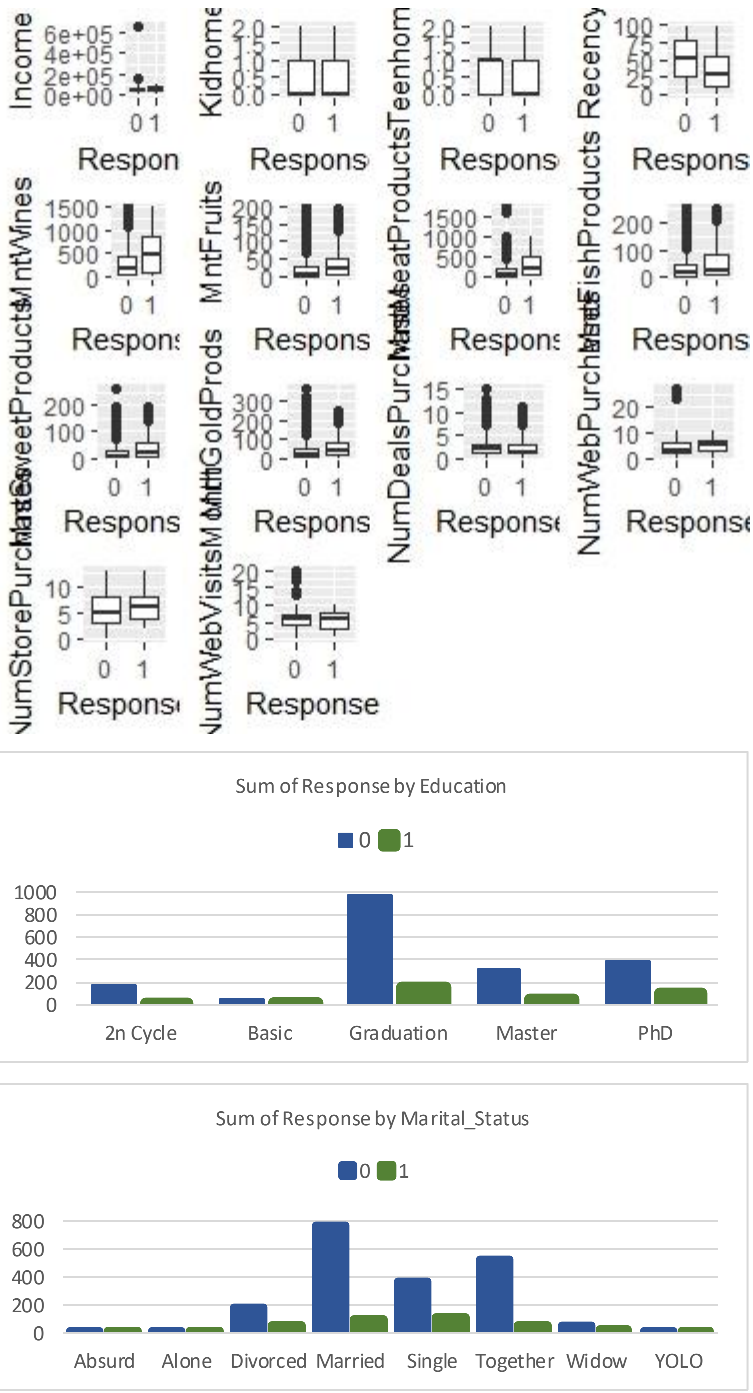

Figure 4 Plot Data

Through this research, it was found that the data set used in the study had poor data quality. In addition to data that has many outliers and missing values, the data is also unbalanced. Unbalanced data causes errors in classification which tend to occur in the minority class [20]. The minority class will be difficult to predict because there is little data on that class [21]. Result of balancing the data with SMOTE are shown in Figure 5.

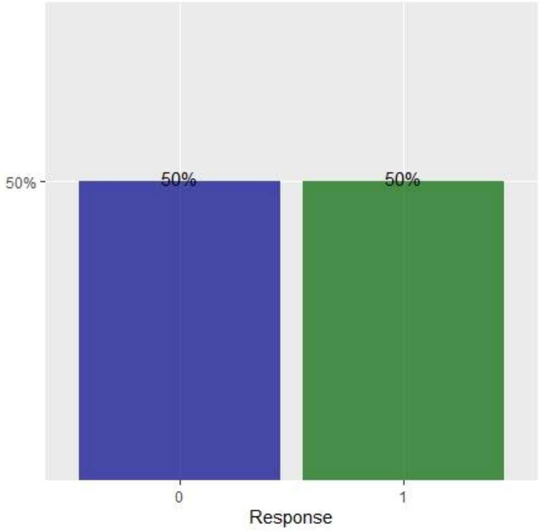

Figure 5 Balanced Data

The evaluation results of the LightGBM model from unbalanced data using the Confussion Matrix show that the accuracy level of classification results is quite good, where the accuracy level of the data train is 86.88% and the accuracy rate of the test data is 85.58%. The predicted results for test data in class 0 or negative responses were 434 customers and class 1 or positive responses were 3 customers. It can also be seen, the results of the classification of test data in Figure 6 of class 0 prices no predictions missed, the results of class 1 price predictions missed as many as 63 customers. This LightGBM method can be said to be Fit because the difference in classification accuracy in train data and test data is quite small, which is 1.3%.

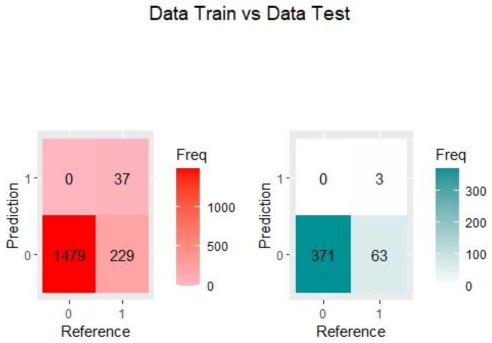

Figure 6 Model LightGBM Evaluation on Data Train vs Data Test Using Confusion Matrix

While the evaluation results of the Random Forest model from balanced data using the Confussion Matrix in Figure 7 show a very good level of accuracy of classification results, where the accuracy level of the data train is 91.53% and the accuracy level of the test data is 99.66%. The prediction results for test data in class 0 or negative responses were 878 customers and class 1 or positive responses were 867 customers. It can also be seen, the results of the classification of test data in Figure 7 at the class 0 price there are data that missed as many as 26 customers, and the prediction results of class 1 prices missed as many as 11 customers [22]. The Random Forest method tends to be underfitting because the difference in classification accuracy in the train data and test data is quite small at 8.13% [23].

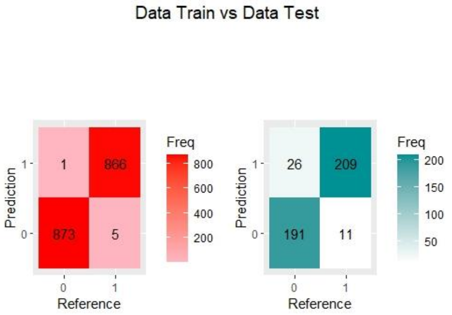

Figure 7 Model Random Forest Evaluation on Data Train vs Data Test Using Confusion Matrix

Overall the accuracy measurement results of the two models are shown in Table 3.

Table 3 Model Accuracy Meas urement

Model		Accuracy	Recall	Specificity	Precision
LightGBM	Imbalance	0.856	0.995	0.0455	0.858
	Balance	0.534	0	1	-
Random Forest with SMOTE	Imbalance	-	-	-	-
	Balance	0.915	0.95	0.880	0.889

Table 3 shows that LightGBM can achieve an accuracy of up to 85.49% on unbalanced data [24]. Even though the Random Forest is not able to form an accurate model on the same data. If the data is balanced using the SMOTE technique, the Random Forest will produce a better accuracy rate of 88.3%, while LightGBM does not affect the results. LightGBM produces a Recall value of up to 99.5% even on unbalanced data. The Specificity value of the Random Forest method has a value of 85.1% while LightGBM is 4.55%. It takes LightGBM method to get the model Faster than Random Forest [25].

4. CONCLUSION

The results of this study is mutually influential variables are Income, MntWines, MntFruits, MntMeatProduct, and NumCatalogPurchase, and NumWebVisitsMonth. There are 3 potential customers to the LightGBM model and there are 234 potential customers according to the Random Forest model from 437 customers.

The result indicate that the LightGBM method has better accuracy in unbalanced data than the Random Forest, which is 85.49%, when the Random Forest is unable to produce a model. The SMOTE method used in this study affects the accuracy of the random forest but does not affect the accuracy of LightGBM. Over all the Accuracy, Recall, Specificity, and Precision values, in Random Forest produce a good value compared to LightGBM on balanced data. Meanwhile, LightGBM can handle unbalanced data and get the model faster than Random Forest.

REFERENCES

[1] P. Cortez and R. Laureano, “Using Data Mining for Bank Direct Marketing : An Application of the CRISP-DM Methodology,” in *European Simulation And Modelling Conference*, 2011.

[2] E. Miranda and Julisar, “DATA MINING DENGAN METODE KLASIFIKASI NAÏVE BAYES UNTUK MENGKLASIFIKASIKAN PELANGGAN Eka Miranda , Julisar Program Sistem Informasi , Program Studi Sistem Informasi , Universitas Bina Nusantara,” *Infotech*, vol. 4, no. 9, pp. 6–12, 2018.

[3] M. E. Lasulika, “Komparasi Naïve Bayes, Support Vector Machine Dan K-Nearest Neighbor Untuk Mengetahui Akurasi Tertinggi Pada Prediksi Kelancaran Pembayaran Tv Kabel,” *Ilk. J. Ilm.*, vol. 11, no.

1, pp. 11–16, 2019, doi: 10.33096/ilkom.v11i1.408.11-16.

[4] D. Mutiara, C. Hermanto, A. F. Sugianto, O. Ika, and A. Nugroho, "Pemilihan Pelanggan Potensial Dengan Melakukan Pemetaan Area Dengan Metode Algoritma K-NN dan K-Means Di Yamaha Nusantara Motor Purwokerto," vol. 4, pp. 52–57, 2021.

[5] R. Siringoringo and I. K. Jaya, "Ensemble Learning Dengan Metode Smote Bagging Pada Klasifikasi Data Tidak Seimbang," vol. 3, no. 2, pp. 75–81, 2018.

[6] L. M. Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," vol. 13, no. 1, pp. 33–44, 2022.

[7] I. Wardana and V. Isnaini, "Gradient Boosting Machine , Random Forest dan Light GBM untuk," no. February, 2022, doi: 10.29207/resti.v5i1.3682.

[8] K. Handayani and Erni, "PENERAPAN LIGHT GRADIENT BOOSTING DALAM PREDIKSI RASIO KLIK," *J. Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 13–18, 2023.

[9] W. Nugraha, "Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi," *J. Manag. dan Inform.*, vol. 9, no. 2, pp. 3–8, 2021.

[10] A. Miftahusalam, A. F. Nuraini, A. A. Khoirunisa, and H. Pratiwi, "Perbandingan Algoritma Random Forest, Naïve Bayes, dan Support Vector Machine Pada Analisis Sentimen Twitter Mengenai Opini Masyarakat Terhadap Penghapusan Tenaga Honorer," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 563–572, 2022, doi: 10.34123/semnasoffstat.v2022i1.1410.

[11] B. Bawono and R. Wasono, "Perbandingan Metode Random Forest dan Naive Bayes," *J. Sains dan Sist. Inf.*, vol. 3, no. 7, pp. 343–348, 2019, [Online]. Available: http://prosiding.unimus.ac.id

[12] Ramadani and B. H. Hayadi, "Perbandingan Metode Naive Bayes Dan Random Forest Untuk Menentukan Prestasi Belajar Siswa Pada Jurusan RPL (Studi Kasus SMK Swasta Siti Banun Sigambal)," *J. Comput. Sci. Inf. Technol. Progr. Stud. Teknol. Inf.*, no. 2, p. 2022, 2022, [Online]. Available: http://jurnal.ulb.ac.id/index.php/JCoInT/index

[13] R. Leonardo, J. Pratama, and Chrisnatalis, "Perbandingan Metode Random Forest Dan Naïve Bayes Dalam Prediksi Keberhasilan Klien Telemarketing," vol. 3, pp. 455–459, 2020.

[14] T. A.M and A. Yaqin, "Perbandingan Algoritma Naïve Bayes, K-Nearest Neighbors dan Random Forest untuk Klasifikasi Sentimen Terhadap BPJS Kesehatan pada Media Twitter," *InComTech J. Telekomun. dan Komput.*, vol. 12, no. 1, p. 01, 2022, doi: 10.22441/incomtech.v12i1.13642.

[15] D. H. Depari *et al.*, "Perbandingan Model Decision Tree , Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," vol. 4221, pp. 239–248, 2022.

[16] L. Sari, A. Romadloni, and R. Listiyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random," vol. 14, no. 01, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.

[17] A. Raza, "Superstore Marketing Campaign Dataset," *kaggle*, 2023. https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset

[18] W. Romadhona, B. I. Nugroho, and A. A. Murtopo, "Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma," vol. 11, no. September, pp. 100–104, 2022.

[19] P. H. Putra, B. Purba, and Y. A. Dalimunthe, "Random forest and decision tree algorithms for car price prediction," vol. 1, no. 2, pp. 81–89, 2023.

[20] N. Putu, Y. Trisna, E. N. Kencana, and I. W. Sumarjaya, "SMOTE : POTENSI DAN KEKURANGANNYA PADA SURVEI," vol. 10, no. November, pp. 235–240, 2021.

[21] Y. . Suhanda, L. . Nurlaela, I. . Kurniati, A. Dharmalau, and I. . Rosita, "Predictive Analysis of Customer Retention Using the Random Forest Algorithm", *TIERS*, vol. 3, no. 1, pp. 35-47, Jun. 2022.

[22] I. G. K. K. . Putra and I. G. W. S. . Dharma, "Application of The K-Means Clustering Method To Search For Potential Tourists of Bendesa Hotel", TIERS, vol. 4, no. 1, pp. 8-15, Jun. 2023.

[23] A. W. O. . Gama, J. T. . Junieargo, and D. A. P. A. G. . Putri, “Rancang Bangun Sistem Informasi Akademik Berbasis Mobile Application”, *TIERS*, vol. 2, no. 1, pp. 31-40, Dec. 2021.

[24] F. D. Marleny and M. Mambang, “Predictive Modeling Classification of Post-Flood and Abrasion Effects With Deep Learning Approach”, *TIERS*, vol. 3, no. 1, pp. 1-10, Jun. 2022.

[25] P. Subarkah, S. A. . Solikhatin, I. . Darmayanti, A. N. . Ikhsan, D. U. . Hidayah, and R. M. . Anjani, “Prediction of Education Level in Population Data Using Naïve Bayes Algorithm”, *TIERS*, vol. 3, no. 2, pp. 69-75, Dec. 2022.