Journal of Socioeconomics and Development

https://publishing-widyagama.ac.id/ejournal-v2/index.php/jsed

Development of direct marketing strategy for banking industry: The use of a Chi-squared Automatic Interaction Detector (CHAID) in deposit subscription classification

Anwar Fitrianto[1]*, Wan Zuki Azman Wan Muhamad[2], and Budi Susetyo[1]

[1]Department of Statistics, IPB University, Bogor, Indonesia
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia

*Correspondence email: anwarstat@gmail.com

ARTICLE INFO

►**Research Article**

Article History
Received 11 January 2022
Accepted 24 February 2022
Published 14 March 2022

Keywords
banking industry; classification; deposit subscriptions; marketing strategy

JEL Classification
A10; C10; G21

ABSTRACT

A comparison between Chi-squared Automatic Interaction Detector (CHAID) and logistic regression analysis was performed for classification problems on bank direct marketing data. CHAID Performance Comparison and comparison with Logistic Regression (LR) performance were also conducted. Priority performance with two statistical measures was evaluated: classification accuracy and sensitivity in the presence of data containing categorical imbalances. Random over sampling (ROS) was then applied to deal with class balance problems to get better performance of CHAID analysis. Segmentation analysis was also performed using the CHAID approach to improve the performance of the analysis results. CHAID outperforms LR because of its advantages that it can be used to perform segmentation modeling. Direct marketers should pay attention to traits are Duration, Month, Contact, and Housing. To get a higher subscription, the bank must extend the call duration. Based on these results, the banking industry needs to prepare regulations related to human resources, infrastructure, costs, and government support to achieve higher subscriptions.

To cite this article: Fitrianto, A., Wan Muhamad, W. Z. A., & Susetyo, B. (2022). Development of direct marketing strategy for banking industry: The use of A Chi-squared Automatic Interaction Detector (CHAID) in deposit subscription classification . Journal of Socioeconomics and Development, 5(1), 64-75. https://doi.org/10.31328/jsed.v5i1.3420

ISSN 2615-6075 online; ISSN 2615-6946 print
©UWG Press, 2022

INTRODUCTION

Classification is one method that is often used in statistical analysis and machine learning. The method, whose process involves data grouping according to similarities, has been widely used in various fields such as e-commerce, direct marketing, market segmentation, environmental research, medical and social sciences, etc. In direct marketing, many classification approaches have been employed to map potential customers so that the bank can determine correct campaign strategies (Amzile & Amzile, 2021; Rogić & Kašćelan, 2021). Another application of the classification approaches in direct marketing is classifying banking products into meaningful categories. This technique helps the bank determine customers' buying behavior (Ładyżyński et al., 2019). In addition, the selection and marketing strategy of an appropriate bank product will help in promoting the services and product of the business.

Data mining is a process for managing data by extracting previously unknown information from large datasets (Edastama et al., 2021). Besides being useful in classification problems, techniques in data mining can be used to see the relationship between a set of independent variables and a dependent variable (Yang

et al., 2020). Many people use multiple linear regression analysis to see the relationship between the independent and dependent variables. Unfortunately, this approach is not suitable when dealing with the categorical dependent variable, such as nominal or ordinal (Klee et al., 2018). When the dependent variable is categorical, certain procedures such as the use of link function are needed to convert the variable into a continuous scale. This is a common problem that practitioners often face. Market researchers often work with cases using categorical variables such as education level, gender, income level, education level, race, etc. A common example might be a situation where a direct marketer who sells newspaper subscriptions wants to maximize his profits by identifying subscribers who are potential household customer segments. The problem of segmentation or classification like this requires a method that is able to perform modeling for segmenting the customers in which the population needs to be divided into different segments regarding some specified criteria (de Caigny et al., 2018).

Data mining is often used in several industries including insurance and banking. In banking industries, data mining techniques in direct marketing aim to analyze customer data and develop techniques to classify customers based on products and services preferred by customers. In solving classification problems, methods or techniques aim to simplify the classification process. Some of the techniques used in problem classification are decision tree, regression tree, deep learning, Suupport Vector Machine, G-Statistics, CHAID, etc. The Support Vector Machine or SVM model is a supervised learning for classification and regression problems (Sen et al., 2020). The SVM is able to solve classification problems for big data, especially in multi-domain application problems in big data (Ghosh & Sanyal, 2018). Meanwhile, AdaBoost model is a popular model in machine learning that is easy to implement and can be applied to recognition and classification problems. However, for classification problems, this model corrects errors made by weak classifiers, making it more prone to overfitting compared to another learning model. A research has been conducted by Siregar et al. (2020) to use a deep learning method to classify customers who have taken these time deposits into marketing targets. This technique has succeeded in classifying potential customers to deposit money in a bank. Meanwhile, Hao et al. (2020) used ensemble learning techniques to classify banks in Portugal for direct marketing planning purposes. CHAID analysis was used by Díaz-Pérez et al. (2020) to construct the visitor's means of transportation according to geographical origin and few other variables such as length of stay, accommodation establishment, season, age, gender, and education level. The research results are helpful for the policy makers.

The problem faced in direct marketing is how to achieve high accuracy in the classification process based on the accuracy of certain information obtained from customers and considered important by banks (Siregar et al., 2020). The main target of bank direct marketing campaigns is to predict consumers' expectations who have the highest possibility in using the bank services using data mining techniques (Fitriani & Febrianto, 2021). The implementation of the in techniques bank direct marketing is used on bank customer loans. Banks must select customers who have the potential to apply for credit.

In data mining, CHAID has been frequently used as one of the classification methods. Besides being used for discovering the relationship between variables, it can be employed as a predictive model to result in conclusions about a target or dependent variable based on a set of predictors (Xu et al., 2019). In the predictive modeling, the objects will split into partitions. By examining each partition individually, the most influential variable in discriminating each partition is identified from the remaining variables. In direct marketing, Veeramuthu (2019) has applied the methods. The performance of CHAID and CART was compared and it was concluded that most of the variables used were categorical. In the comparison, the study found that CHAID had better performance. Meanwhile, the study also found that if most of the variables used are continuous scale, it is better to use CART. Furthermore, Díaz-Pérez et al. (2020) conducted a study showing that CHAID can be applied to study the segmentation of tourists. In that case, CHAID works based on the tourist's likelihood of revisiting a tourism destination. Using CHAID, they were able to carry out a characteristics identification of travel in each segment, which is very useful for preparing strategies in tourism marketing.

Classification in data mining has problems when the dataset is imbalanced. The imbalanced dataset is a situation when the distribution of the category of interest in the dataset is not approximately equally. Marinakos & Daskalaki (2017) conducted a study

about the problem of classification for an imbalanced dataset. In the paper, comparisons of performance among statistical methods, induction, Machine Learning, and distance-based classification algorithms were conducted. The research aimed to predict potential depositors when the dataset is imbalanced. The concern was the effort to effectively balance the data set during training for the negative effects of imbalance and to improve the correct classification of underrepresented classes.

A research conducted by Kaur et al. (2019) applied few methods, i.e. oversampling, undersampling, boosting, and bagging to determine which method has better performance for handling imbalanced datasets. The predictive performance of CHAID, CART and C5 was thoroughly assessed after resampling approach. In this case, resampling needed to be done to reduce the effect of category imbalance. Two resampling techniques that are often used are random oversampling (ROS) and random undersampling (RUS). In ROS, the technique randomly duplicates the minority class sample to modify the class distribution. Unfortunately, this technique can cause overfitting. Meanwhile, RUS works by discarding the primary random sample, which will also discard the potentially useful sample. When dealing with imbalanced data sets, the performance of the classification technique can be measured in terms of precision and sensitivity. This is because the level of accuracy of the predictive model is adjusted to the size. If this is not done, it will result in a bias towards the favored class. A study by Møller et al. (2019) found that implementing a random resampling strategy improves classification performance for imbalanced data sets using decision trees than applying bagging and boosting.

Banking is a sector that produces data and transactions recorded every day. One of the daily information recorded in the banking industry is the subscription bank deposit. This data need to be classified to expand the business by improving marketing campaigns. This type of data always grows from time to time. With the increasing data, banking institutions will have difficulty predicting whether their customers will subscribe to term deposits or not, so the banking institution must do the most appropriate mechanism to classify the data. One of the classification techniques that can be used is CHAID.

The study aims to conduct classification modeling for bank deposit subscriptions by comparing Logistic Regression (LR) and CHAID classification tree to determine the better method to describe the bank's direct marketing data. One of the direct marketing data that the bank industries need to classify was the bank deposits subscriptions. The logistic regression analysis and CHAID will be used since the two methods have similar functions in classification.

RESEARCH METHOD

Data Description

In this study, secondary data taken from the Kaggle warehouse were used. The data were about direct marketing conducted by banks in Portugal acquired from 41,188 observations consisting of 20 attributes and 2 classes. This dataset were related to direct marketing carried out by Portuguese banking institutions to their potential customers. The marketing campaign was carried using phone calls. Typically, it takes more than one bidding process to the same prospective customer to determine whether to use the products offered or not. The corresponding information about the data is displayed in Table 1.

The potential customers or clients were asked more than once to determine whether the product (bank time deposit) was subscribed. In this study, the dependent variable is ordering bank time deposit, which was coded as 1 if the customer ordered bank deposits and 0 otherwise. There are 17 independent variables consisting of categorical and continuous variables, i.e. Marital status, Age, Education level, Job, Housing, Credit, Balance, Housing, Contact, Day, Month, Duration, Campaign, Day, Previous, Pdays, and Poutcome.

The data were then analyzed using the steps as in Figure 1. The analysis involved specifying independent and dependent variables, comparing logistic regression and the CHAID analysis, checking whether the data were balanced or imbalanced, and conducting more detailed analysis including segmentation of the clients.

CHAID Algorithm

CHAID is a classification technique that uses the Chi-square test as its primary tool. In general, the technique works by studying the relationship between the dependent variable and several independent variables. It is an iterative technique that examines the independent variables used in the classification one-by-one and arranges them based on the level of significance of the independent variables. The

technique works by splitting the dependent variables into two or more categories using regression or decision tree approaches. The selection of the independent variables that significantly affect the classification is based on the Chi-square test. It is a non-parametric test that does not require assumptions and is suitable for testing the relationship between categorical variables.

Table 1. Descriptions of the Data

No.	Abbreviation	Definition	Data Types
Bank client data			
1	Age	Client's age	Continuous
2	Marital	Marital status	Nominal
3	Job	Client's job type	Nominal
4	Education	Client's education level	Ordinal
5	Balance	Average yearly balance, in euros	Continuous
6	Credit	Status, whether the client has credit in default	Binary
7	Housing	Status, whether the client owns housing loan	Binary
8	Loan	Does client has personal loan?	Binary
Items related to current campaign's last contact to the client			
9	Contact	Contact communication type	Nominal
10	Month	Last contact to the client (name of month in that year)	Nominal
11	Day	Last contact to the client (ith day of the month in that year)	Nominal
12	Duration	The duration of the last contact (in second)	Continuous
Other attributes			
13	Campaign	Number of contacts performed to the clients, calculated from the first to the last contact)	Continuous
14	Previous	Number of contacts performed before the campaign and for this client	Continuous
15	Pdays	Number of days which is calculated after the last contact to the client from the previous campaign (it equals -1 when the client was not contacted, previously)	Continuous
16	Poutcome	Outcome obtained from the previous marketing campaign	Ordinal

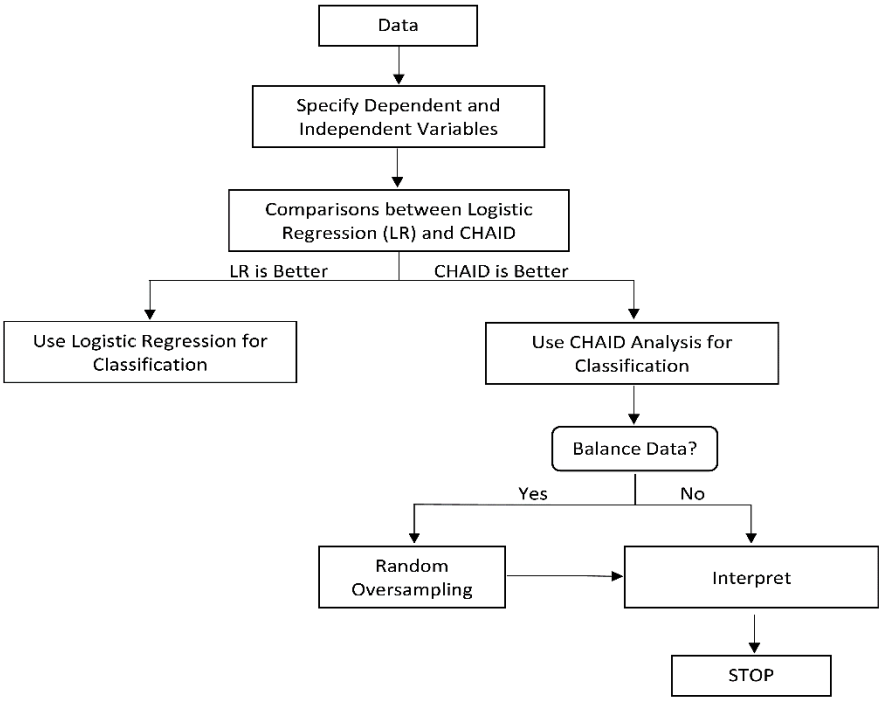

Figure 1. Data analysis step of the study

The classification using CHAID mainly consists of merging, splitting, and stopping conducted repeatedly on each node. The algorithm of CHAID follows the steps mentioned in Aksoy et al. (2018), Bethencourt-Cejas & Diaz-Perez (2017), and Díaz-Pérez et al. (2020). Before carrying out the first step, the continuous predictor must be converted into a categorical predictor variable. Meanwhile, the categorical predictors have categories that were defined naturally. Then, the continuous variables were distributed approximately equally into the categories. The analysis of CHAID method was presented in the form of a tree diagram, making it more exciting and easier to interpret.

Merging was the beginning step of CHAID algorithm. This step merged categorical pairs of each independent variable that provide the most negligible contribution to the dependent variable. The allowed categorization varied according to the type of independent variable. For example, a floating predictor category and a monotonous predictor could only be combined with adjacent categories, while the independent predictors could be combined in any way.

The second step in CHAID was to select the split variable. Independent variables with adjusted p-values in various dependent variables (target variables) were selected to produce the most significant separation (Milanović & Stamenković, 2016). It could be evaluated based on Bonferroni-adjusted p-value. When the p-value of a selected predictor variable falls below the specified value of alpha-to-split, the predictor was selected. In that case, the tree used the independent variable to split with the merged categories. No further splits would be performed if the further adjusted value of the Bonferroni predictor was greater than the specified value of alpha-to-split. In that stage, the process of splitting ended, and each node would become a terminal node.

Merging and splitting steps were repeated on each new node until the specified stopping criteria were met. Research by McCordic & Frayne (2017) discussed a few of the stopping criteria as follows: (i) maximum level number of the tree, (ii) the threshold of the alpha-to-split, (iii) the size of the minimal parent node (the split would not be performed on nodes having smaller than min_{parent} cases), and (iv) the size of the minimal node (The performed split resulted in categories with at least min_{node} each)

Chi-Square Test

The Chi-square test is one of the methods in non-parametric statistical analysis commonly used to examine the association between two categorical variables or for testing the goodness-of-fit of a data set in which the data consist of frequencies. In this study, to examine whether the categorical variables are associated with each other, the Chi-square test was used. The calculation of the test was done by pairing the categorical variables, and the frequencies were displayed in a contingency table (Everitt, 2019). Frequencies of combination between categories were placed in a contingency table cell. The frequency was referred to as observed frequency. The Chi-square statistic is written as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

where E_i and O_i are the expected and observed frequency of the contingency table, respectively. Referring to Equation 1, the Chi-square statistics were calculated by taking the difference between the observed and expected frequencies of each combination of categories of the variables. This difference was then squared and divided by the expected frequency in that category. The summation of the values were taken for all the categories and the result of the summation was referred to as the Chi-squared value. The Chi-square value was compared with Chi-square table with the corresponding degree of freedom of:

$$d.f. = (r-1)(c-1) \quad (2)$$

where r, c are the number of rows and columns, respectively.

Random Oversampling

In classification techniques, there will be problems when there is a class imbalance due to a situation when the size of the minority class is small, and it is not nearly equal to the class majority. To resolve the problem, certain sampling technique needs to be employed to change the minority class distribution to avoid less representative sample training data. To solve the oversampling problem, a resampling technique needs to be conducted for changing the distribution of the minority class so that underrepresented sample could be prevented. According to Xie et al. (2019), the imbalanced class problem can be solved by conducting oversampling to

the class minority. It is conducted by randomly replicating the existing minority class until the number of samples in the majority class is close to the number of samples in the minority class. When the distribution of minority class and majority class is balanced, then the sample in each makes it possible to train the model to recognize that class accurately. The advantage of doing the oversampling is that it can prevent the selection of samples that may be useful, but it can result in overfitting and produce models that are too complex.

Classification Evaluation

According to Olosunde & Soyinkab (2020), the error rate, also known as the probability of misclassification, needs to be calculated to see the performance of the classification model. A performance measure that can be used is to calculate the apparent error rate (AER), which is formulated as the fraction of observations in a sample that are misclassified by a classification model (Santos et al., 2020). Besides AER, the performance of a classification model can be evaluated based on another measure. The standard performance measurement for classification problems that is commonly used is accuracy. However, for cases with an imbalanced class distribution, it is necessary to use specificity, sensitivity, and precision as additional measurements (Thabtah et al., 2020).

Table 2. Confusion Matrix for Evaluation Classification

Actual	Predicted	
	Negative (No)	Positive (Yes)
Negative (No)	TN	FP
Positive (Yes)	FN	TP

Accuracy is the ability of the model to perform the classification process correctly. Whereas sensitivity refers to correctly classified positives, specificity refers to correctly classified negatives, and precision is the positive predictive value defined as the ratio of correctly predicted entities to the total number of entities classified by the model. The number of entities each group (class) involved in the calculation that is classified correctly will be represented in a confusion matrix. It contains information about actual and predicted classifications which is conducted by the classification model (Hasnain et al., 2020). The structure of the confusion matrix is displayed in Table 2.

Confusion matrix is a method that is usually used to perform the accuracy measures on the concept of data mining. There are four terms to represent the classification results. The four terms are True Positive (TP) which is a positive value that is detected correctly, True Negative (TN) is the number of negative data that is detected correctly, False Positive (FP) is negative data but detected positively and False Negative (FN) is negative data is detected as negative data. The detail of the measures of classification performances are as follows:

The calculation to obtain the apparent error rate (AER) as follows:

$$AER = \frac{FN + FP}{TFN + FP + TP + TN} \quad (3)$$

The predictive model has a measure of accuracy which is formulated as:

$$accuracy = \frac{TP + TN}{FN + FP + TP + TN} \quad (4)$$

The predictive model has a measure of sensitivity which is written as:

$$sensitivity = \frac{TP}{FN + TP} \quad (5)$$

The specificity of the predictive model is computed as:

$$specificity = \frac{TN}{FP + TN} \quad (6)$$

And the model precision is computed as :

$$precision = \frac{TP}{FP + TP} \quad (7)$$

RESULT AND DISCUSSION

CHAID Classification Tree

CHAID analysis is available in various statistical software. Regardless of the software used, there is always a procedure for removing independent variables that do not make a major contribution to the classification modeling. SPSS Modeler version 14.1, which was used to perform the CHAID analysis in this study, automatically excluded independent variables that did not make a significant contribution.

The results of the analysis show that there were 9 out of 16 independent variables used in this study that made a significant contribution to this classification. The variables were Age, Housing, Contact, Day, Marriage, Month, Duration, Previous, Duration, and Poutcome. Meanwhile, the independent variables that did not make a significant contribution to the

formation of the best classification model were Education, Loan, Job, Balance, Default, Days, and Campaign. The default value used for tree depth in the SPSS Modeler in this study was 5. However, to obtain a less complicated classification tree, a tree depth of 4 was used. Meanwhile, the minimum case values in the child and parent nodes used in this study were 0.5% and 0.1%, respectively. Then, the data were divided into training and testing set data with a composition of 70% and 30%. With this composition, the resulting classification tree had 82 terminal nodes and 126 total nodes. The total number of nodes and terminals produced a confusion matrix, and the CHAID classification tree measures of performance are shown in Table 3 and Table 4.

Table 3. Confusion Matrix of the CHAID Analysis

Actual	Predicted			
	Training set		Testing set	
	Negative (No)	Positive (Yes)	Negative (No)	Positive (Yes)
Negative (No)	19142	567	19585	628
Positive (Yes)	1681	958	1623	1027

Table 4. Performance Measure of the CHAID Analysis

	Training set	Testing set
AER	0.1006	0.0985
Accuracy	0.8994	0.9015
Sensitivity	0.3630	0.3875
Specificity	0.9712	0.9689
Precision	0.6282	0.6205

Comparison between CHAID and Logistic Regression

Several machine learning algorithms and logistic regression have been widely used for research in the banking world. An example is to find an effective search in direct bank marketing. In this study, the performance between logistic regression and CHAID was compared based on several measures, the results of which are presented in Table 4 and Table 5. There was a class imbalance in this study where the ratio between the minority class to the majority class was 13:100. As previously explained, imbalanced data conditions will result in low accuracy of the prediction model, leading to bias in the majority class. This condition encouraged the measurement of the performance of the classification model using sensitivity and precision measures.

Based on the analysis, in developing the direct marketing model, both logistic regression and CHAID gave similar performance (Table 5). Logistic regression showed higher precision, but CHAID performance had higher sensitivity. Meanwhile, logistic regression had a higher AER (Apparent Error Rate) value than CHAID. The situation indicates that the misclassified observations for training and testing assigned by logistic regression analysis are higher. Moreover, in segmenting clients, logistic regression analysis failed to perform the task of being a homogeneous group and also to characterize or differentiate each segment. In this case, CHAID performs better. The direct bank marketing development strategy will be greatly helped by analyzing the character for each segment. With the results of the study mentioned at the beginning of this paragraph, it shows that the analysis using CHAID produces better information in relation to direct bank marketing. In addition, due to imbalanced data, the low value of model sensitivity indicates that more than half of potential clients are likely to lose.

Table 5. Comparison Performance between Logistic Regression and CHAID

Algorithm	Error on Training Set	Error on Testing Set	Sensi-tivity	Preci-sion
Logistic Regression	0.1012	0.0993	0.3170	0.6430
CHAID	0.1006	0.0985	0.3630	0.6282

Handling the Imbalance Data

Random oversampling is often used to solve the problem of imbalanced data in order to improve classification performance. In this study SPSS Modeler 14.1 has been used to do this. The thing that needs to be known before doing random oversampling is the distribution of each class of the dependent variable. After getting this information, then random oversampling was implemented by replicating the minority class so that the distribution is more or less the same as the majority class. Implementing the random oversampling produced training data in which the distribution of the minority class to the majority class in the training data had a more balanced ratio, which was 19940:19709. The next analysis implemented CHAID approach on the balanced data to train the classification model. Table 6 and Table 7 present the results of this approach in the form of a Confusion Matrix and several classification performance measures.

It can be seen from Table 7 that with the implementation of random oversampling, in terms of sensitivity and precision, there was an increase in

CHAID performance in the training data. Unfortunately, there had been an increase in the number of false positives, reflecting the poor performance of this CHAID approach in terms of precision on the training data. Due to the model with the desired sensitivity, the number of false predictions would be increased to obtain the most. For this reason, the number of false predictions must be increased in order to obtain the desired condition, which was a high sensitivity value. One way that needed to be done was to segment with the CHAID approach on data already balanced by oversampling.

Table 6. Confusion Matrix of CHAID Classification Tree on Balanced Dataset

Actual	Predicted			
	Training set		Testing set	
	Negative (No)	Positive (Yes)	Negative (No)	Positive (Yes)
Negative (No)	15386	4323	15649	4564
Positive (Yes)	3349	16591	446	2204

Table 7. Performance Measure of CHAID on Balanced Dataset

	Training data	Test data
AER	0.1935	0.2191
Accuracy	0.8065	0.7809
Sensitivity	0.8320	0.8317
Specificity	0.7807	0.7742
Precision	0.7933	0.3257

Of the 16 independent variables used as the basis for classifying clients, after applying random oversampling, 11 independent variables were found as the important variables in carrying out this classification, i.e. Marital Status, Age, Housing, Campaign, Balance, Duration, Day, Month, Contact, Poutcome and Days. Meanwhile, the independent variables that did not contribute significantly to the client classification which were then excluded from the model were Education, Loans, Previous, Default and Employment. In terms of the depth of the classification tree, as mentioned in the previous section, this study used deep tree 4 to reduce complexity. With the increase in the number of certain cases in the training data after implementing random oversampling, the minimum cases for the parent and child nodes were set at 1% and 2%, respectively. In addition, with the proportion of training data and test data with a ratio of 70% and 30%, the total number of nodes was 94 and terminal nodes were 63 in the classification tree. Each sample sub-group was represented by one node in the classification tree.

Analysis of Improved CHAID Classification Tree

The independent variable Duration was the first best variable used to divide clients on the classification tree. Based on this variable, 10 splits were formed as the first split referred to as the parent nodes. With a depth of value of 4, each node would be split into nodes up to a maximum of 4 levels. The node at the lowest level was the root node, with the top node containing all samples and the absolute frequency of each category of independent variables. The proportion and frequency of each client category of independent variables would be contained in each node in the classification tree. Whereas, the terminal node was the last segment that was not further divided. The result of the entire process in this study was the prediction of the percentage of clients in each segment making subscription and non-subscription for the bank term deposit, as presented in Table 8.

Table 8 presents the number of subscripts and the percentage of each segment formed in the classification tree. Of the 63 segments formed, Segment 39 is the segment that has the largest percentage of subscriptions (95.66%). This study found that contacts made with clients were between 259 and 325 seconds, that is, the last contacts were in March, September, October and December. Both information indicates that a high number of subscriptions could be obtained when campaign was launched in March, September, October or December, with a call duration to the client for more than 259, but not more than 325 seconds. The five segments with the next highest subscriptions in a row from high to low were Segment 50 (94.20%), Segment 61 (93.83%), Segment 58 (93.76%), Segment 22 (93.32 %), and Segment 31 (92.92%). In general, Table 8 also provides information that there were 19 segments (31.15%) with a subscription percentage of more than 75%. In general, segments with high percentages values are symbolized by bigger size bubbles (Figure 2). These segments should be used as marketing targets for banking institutions. The characteristics of these segments should be studied in more detail because they are very potential customers.

Table 8. Segmentation of Subscription of the Bank Term Deposit

Segment	Subscription		Non-Subscription		Segment	Subscription		Non-Subscription	
	Frequency	Proportion %	Frequency	Proportion %		Frequency	Proportion %	Frequency	Proportion %
1	115	19.56	473	80.44	33	74	18.55	325	81.45
2	42	9.42	404	90.58	34	8	1.70	464	98.31
3	15	2.91	501	97.09	35	516	67.81	245	32.19
4	0	0.00	634	100.00	36	310	39.29	479	60.71
5	0	0.00	1219	100.00	37	379	88.14	51	11.86
6	7	1.18	588	98.82	38	621	94.67	35	5.34
7	501	53.58	434	46.42	39	181	40.13	270	59.87
8	55	7.23	706	92.77	40	194	41.81	270	58.19
9	0	0.00	562	100.00	41	0	0.00	412	100.00
10	14	3.02	450	96.98	42	499	70.18	212	29.82
11	45	10.90	368	89.10	43	537	69.38	237	30.62
12	125	16.19	647	83.81	44	212	48.85	222	51.15
13	451	61.44	283	38.56	45	658	92.68	52	7.32
14	123	30.22	284	69.78	46	368	50.34	363	49.66
15	57	12.13	413	87.87	47	45	7.48	557	92.53
16	383	83.81	74	16.19	48	796	81.73	178	18.28
17	164	28.57	410	71.43	49	268	61.61	167	38.39
18	22	4.97	421	95.03	50	487	94.20	30	5.80
19	0	0.00	403	100.00	51	426	64.74	232	35.26
20	21	4.01	503	95.99	52	559	79.97	140	20.03
21	325	46.90	368	53.10	53	301	45.40	368	54.60
22	447	93.32	32	6.68	54	464	84.98	82	15.02
23	383	66.38	194	33.62	55	433	76.10	136	23.90
24	91	12.22	654	87.79	56	388	87.59	55	12.42
25	45	5.78	733	94.22	57	838	89.15	102	10.85
26	187	28.77	463	71.23	58	436	93.76	29	6.24
27	413	62.39	249	37.61	59	362	81.72	81	18.28
28	159	30.58	361	69.42	60	382	68.58	175	31.42
29	173	43.47	225	56.53	61	1825	93.83	120	6.17
30	347	80.89	82	19.11	62	456	85.55	77	14.45
31	604	92.92	46	7.08	63	1349	91.21	130	8.79
32	216	48.54	229	51.46					

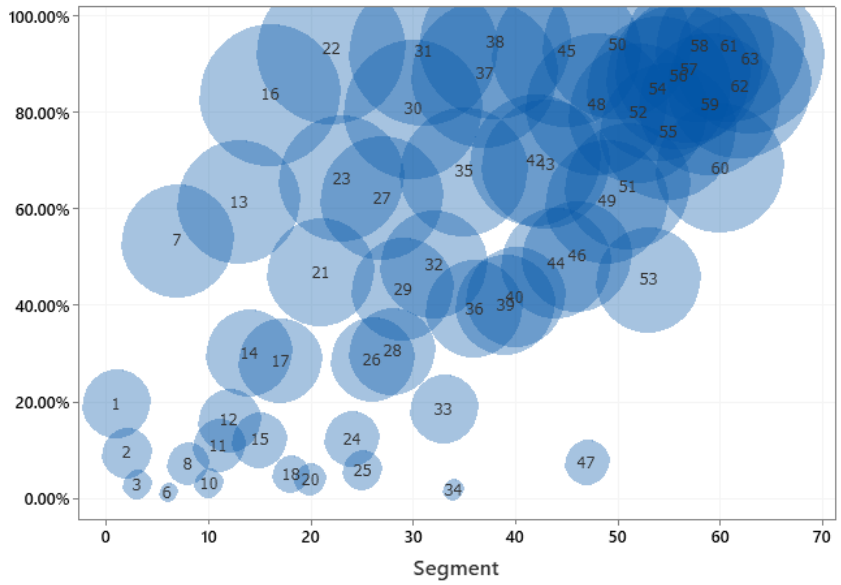

Figure 2. Segmentation of potential bank customers as the size of the bubbles

Research Implication

Data mining, such as classification modeling, is a strategically important area for the banking sector. The research by Idrees (2019) has also examined the use of CHAID as a classification technique for bank deposit subscriptions. In the subscription to bank deposits carried out in this study, it was found that the duration of calls to clients is an important variable to be used to classify clients. More specifically, this study has found that to get a positive response from clients during a campaign, the call duration should not be too short. The longer the call duration, the client will tend to give a positive response. This is in line with a study conducted by Asare-Frempong & Jayabalan (2017). In addition, in order to have a more successful campaign in capturing more customers, direct marketers should also consider contacting the bank's clients either by telephone or cell phone. This is in accordance with what was done by Alexandra & Sinaga (2021). Clients who do not have credit for clients who positively respond to the campaign should be the target of future campaigns (Alexandra & Sinaga, 2021). In order to achieve higher subscriptions, the banking industry needs to prepare several things, such as human resources, infrastructure, costs, and government support in terms of regulation. For human resources, banks need to prepare people who have good communication skills so that customers will feel comfortable receiving phone calls longer than usual. Moreover, the banking industry needs to provide the infrastructure that makes direct marketers comfortable, such as telecommunications equipment. From the government's point of view, the regulations need to support the banking industry to achieve higher subscription rates include lowering interest rates.

Based on the results, in order to achieve a higher subscription, the banking industry needs to prepare regulations related to human resources, infrastructure, costs, and government support. For human resources, banks need to prepare people who have good communication skills so that customers will feel comfortable receiving phone calls for a longer time. For infrastructure, the banking industry needs to provide the infrastructure that makes direct marketers comfortable, such as telecommunications equipment. From the government's perspective, the regulations needs to support the banking industry to achieve a higher subscription include lowering interest rates.

CONCLUSION AND SUGGESTION

This study examines the use of CHAID to classify bank subscription deposits which are useful for segmenting and identifying characteristics that influence customers to subscribe to bank deposits. The classification is based on direct bank marketing data. Before deciding to use CHAID in a more detailed analysis, a comparison was made between logistic regression and CHAID. The result is that CHAID has a better performance than logistic regression. Then a more detailed analysis for segmenting and classifying customers was carried out using CHAID. Because the direct bank marketing data used is imbalanced, a random oversampling technique was used to balance the ratio of minority and minority classes from the training data. The purpose is that the classification model formed can identify minority classes more accurately. The performance of the classification model as measured by the value of sensitivity and precision seemed to improve on the results of the analysis using the CHAID method on balanced training data than on imbalanced training data. The results of the classification modeling formed can be used to target customers who have the potential to subscribe to bank deposits. Based on the obtained classification model, the criteria for potential customers that should be considered include the Housing, Contact, Month, and Duration variables. The results imply that it is recommended targeting clients who do not own any housing loans. Moreover, to capture a higher number of bank deposit subscriptions, the marketing campaign should be launched in March, September, October, and December. Besides that, higher number of bank deposit subscription can be obtained if the direct marketer contacts the clients through mobile phone or telephone at a slightly longer call duration.

REFERENCES

Aksoy, A., Ertürk, Y. E., Eyduran, E., & Tariq, M. M. (2018). Comparing predictive performances of MARS and CHAID algorithms for defining factors affecting final fattening live weight in cultural beef cattle enterprises. Pakistan Journal of Zoology, 50(6), 2279–2286.

https://doi.org/10.17582/journal.pjz/2018.50.6.2279.2286

Alexandra, J., & Sinaga, K. P. (2021). Machine learning approaches for marketing campaign in portuguese banks. 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS), 1–6.

Amzile, K., & Amzile, R. (2021). Using SVM for Smart Direct Marketing (SDM): A case of predicting bank customers interested in the term deposits. International Journal of Accounting, Finance, Auditing, Management and Economics, 2(5), 525–537. https://www.ijafame.org/index.php/ijafame/article/download/366/294/

Asare-Frempong, J., & Jayabalan, M. (2017). Predicting customer response to bank direct telemarketing campaign. In 2017 International Conference on Engineering Technology and Technopreneurship, ICE2T 2017 (Vol. 2017-January, pp. 1–4). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICE2T.2017.8215961

Bethencourt-Cejas, M., & Diaz-Perez, F. M. (2017). An application of the CHAID algorithm to study the environmental impact of visitors to the Teide National Park in Tenerife, Spain. International Business Research, 10(7), 168–177. https://doi.org/10.5539/ibr.v10n7p168

de Caigny, A., Coussement, K., & de Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

Díaz-Pérez, F. M., García-González, C. G., & Fyall, A. (2020). The use of the CHAID algorithm for determining tourism segmentation: A purposeful outcome. Heliyon, 6(7), e04256. https://doi.org/10.1016/j.heliyon.2020.e04256

Edastama, P., Bist, A. S., & Prambudi, A. (2021). Implementation of data mining on glasses sales using the apriori algorithm. International Journal of Cyber and IT Service Management, 1(2), 159–172. Retrieved from https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/46

Everitt, B. S. (2019). The analysis of Contingency Tables. Chapman and Hall/CRC. https://books.google.co.id/books?id=aSe1LbYz3v0C

Fitriani, M. A., & Febrianto, D. C. (2021). Data mining for potential customer segmentation in the marketing bank dataset. JUITA: Jurnal Informatika, 9(1), 25–32. https://doi.org/10.30595/juita.v9i1.7983

Ghosh, M., & Sanyal, G. (2018). An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning. Journal of Big Data, 5(1), 1–25. https://doi.org/10.1186/s40537-018-0152-5

Hao, R., Xia, X., Shen, S., & Yang, X. (2020). Bank direct marketing analysis based on ensemble learning. Journal of Physics: Conference Series, 1627(1), 012026. https://doi.org/10.1088/1742-6596/1627/1/012026

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating trust prediction and confusion matrix measures for web services ranking. IEEE Access, 8, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

Idrees, H. I. H. (2019). Building a Classification and Forecasting Model to Support Decision Making. PhD Thesis, Sudan University of Science and Technology. Retrieved from http://41.67.53.40/handle/123456789/22745?show=full

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys (CSUR), 52(4), 1–36. https://doi.org/10.1145/3343440

Klee, C. A., Rogers, B. M. A., Caravedo, R., & Dietz, L. (2018). Measuring/s/variation among younger generations in a migrant settlement in Lima, Peru. Studies in Hispanic and Lusophone Linguistics, 11(1), 29–57. https://doi.org/10.1515/shll-2018-0002

Ładyżyński, P., Żbikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. Expert Systems with Applications, 134, 28–35. https://doi.org/10.1016/j.eswa.2019.05.020

Marinakos, G., & Daskalaki, S. (2017). Imbalanced customer classification for bank direct marketing. Journal of Marketing Analytics, 5(1), 14–30. https://doi.org/10.1057/s41270-017-0013-7

McCordic, C., & Frayne, B. (2017). Household vulnerability to food price increases: The 2008 crisis in urban Southern Africa. Geographical Research, 55(2), 166–179. https://doi.org/10.1111/1745-5871.12222

Milanović, M., & Stamenković, M. (2016). CHAID decision tree: Methodological frame and

application. Economic Themes, 54(4), 563–586. https://doi.org/10.1515/ethemes-2016-0029

Møller, A. B., Iversen, B. v, Beucher, A., & Greve, M. H. (2019). Prediction of soil drainage classes in Denmark by means of decision tree classification. Geoderma, 352, 314–329. https://doi.org/10.1016/j.geoderma.2017.10.015

Olosunde, A. A., & Soyinkab, A. T. (2020). Discrimination and classification model from multivariate exponential power distribution. Electronic Journal of Applied Statistical Analysis, 13(2), 284–292. https://doi.org/10.1285/i20705948v13n2p284

Rogić, S., & Kašćelan, L. (2021). Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning. Computer Science and Information Systems, 18(3), 893–925. https://doi.org/10.2298/CSIS200530052R

Santos, A. E. M., Amaral, T. K. M., Mendonça, G. A., & Silva, D. de F. S. da. (2020). Open stope stability assessment through artificial intelligence. REM-International Engineering Journal, 73, 395–401. https://doi.org/10.1590/0370-44672020730012

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modelling and graphics (pp. 99–111). Springer. https://doi.org/10.1007/978-981-13-7403-6_11

Siregar, A. M., Faisala, S., Handayania, H. H., & Jalaludinb, A. (2020). Classification data for direct marketing using deep learning. Scientific Journal of PPI-UKM, 7(2), 33–37. Retrieved from http://www.kemalapublisher.com/index.php/ppi-ukm/article/view/456

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: experimental evaluation. Information Sciences, 513, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Veeramuthu, P. (2019). Analysis of progressive duplicate data detection. Journal of Computational Mathematica, 3(2), 41–50. https://doi.org/10.26524/cm53

Xie, W., Liang, G., Dong, Z., Tan, B., & Zhang, B. (2019). An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data. Mathematical Problems in Engineering, 2019, 1–13. https://doi.org/10.1155/2019/3526539

Xu, H., Zhou, J., G Asteris, P., Jahed Armaghani, D., & Tahir, M. M. (2019). Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate. Applied Sciences, 9(18), 3715. https://doi.org/10.3390/app9183715

Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, S., Xu, A., & Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. Journal of Evidence-Based Medicine, 13(1), 57–69. https://doi.org/10.1111/jebm.12373