BOHR International Journal of Smart Computing and Information Technology
2023, Vol. 4, No. 1, pp. 24–38
DOI: 10.54646/bijscit.2023.34
www.bohrpub.com

METHODS

Synthetic feature generation to improve accuracy in prediction of credit limits

Sikha Bagui* **and Jennifer Walker**

Department of Computer Science, University of West Florida, Pensacola, FL, United States

***Correspondence:**
Sikha Bagui,
bagui@uwf.edu

Received: 05 April 2023; **Accepted:** 18 May 2023; **Published:** 06 June 2023

Financial institutions use various data mining algorithms to determine the credit limits for individuals using features like age, education, employment, gender, income, and marital status. But, there is still a question of accurate predictability, that is, how accurate can an institution be in predicting risk and granting credit levels. If an institution grants too low of a credit limit/loan for an individual, then the institution may lose business to competitors, but if the institution grants too high of a credit limit/loan, then the institution may lose money if that individual does not repay the credit/loan. The novelty of this work is that it shows how to improve the accuracy in predicting credit limits/loan amounts using synthetic feature generation. By creating secondary groupings and including both the original binning and the synthetic bins, the classification accuracy and other statistical measures like precision and ROC improved substantially. Hence, our research showed that without synthetic feature generation, the classification rates were low, and the use of synthetic features greatly improved the classification accuracy and other statistical measures.

Keywords: synthetic feature generation, random forest, random tree, REPTree, Naïve Bayes, credit amount, credit risk

1. Introduction

In today's world, there is no question that almost everyone has a credit card. One may need to be 18 years of age before applying for credit, but parents can actually add their young children as authorized users to their credit cards. So the real question asked by banks and credit card companies is whether the primary account holder is at high risk or low risk in repaying the credit or loan, and thereby credit is granted based on risk levels; high risk equals low credit limit, and low risk equals high credit limit. Institutions use various data mining algorithms to determine the credit limits for individuals. The typical features used by institutions to help determine credit limits include age, education, employment, gender, income, and marital status (1). Using these features, there is still a question of accurate predictability, that is, how accurate can an institution be in predicting the risk and granting credit levels. If an institution grants too low of a credit limit/loan for an individual, then the institution may lose business to competitors, but if the institution grants too high of a credit limit/loan, then the institution may lose money if that individual does not repay the credit/loan.

The novelty of this work it that it shows how to improve the accuracy in predicting credit limits/loan amounts using synthetic feature (SF) generation. By creating secondary groupings and including both the original binning and synthetic bins, the classification accuracy and other statistical measures like precision and ROC improved substantially.

Four different datasets were used for this analysis. Three datasets (datasets 1, 3, and 4) were used for prediction of credit limits/loans, and one dataset (dataset 2) was used for predicting bank approvals of credit/loans. In this work, first, feature selection was performed using information gain. Then, the classification was performed using features with higher information gain and the synthetic features. For classification, three different tree-based classifiers (Random

Forest, Random Tree, and REPTree) and one non-tree-based classifier (Naïve Bayes) were used.

The rest of this article is organized as follows. Section 2 presents the related works; section 3 presents the datasets and preprocessing; section 4 presents the results and discussion; and section 5 presents the conclusions.

2. Related works

Zeng (2) studied effective binning on credit scoring. He focused on the weight of the evidence and regression modeling for binning of continuous variables. The feature, age (typically a variable in any financial dataset) was the example used for improving the binning process.

Danenas and Garsva (3) presented their work on credit risk evaluation based on linear support vector machine classifiers. This was combined with external evaluation and testing sliding windows, with a focus on larger dataset applications. These authors concluded that, using real-world financial datasets, for example, from the SEC EDGAR database, their method produced results comparable to other classifiers such as logistic regression and thus could be used for the future development of real credit risk evaluation models.

Lessmann et al. (4) compared several classification algorithms to credit scoring. They examined the extent to which the assessment of alternative scorecards differs across established and novel indicators of predictive accuracy.

Ala'raj and Abbod (5) presented a new ensemble combination approach based on classifier consensus to combine multiple classifier systems of different classification algorithms. Specifically, five well known base classifiers were used: Neural Networks, Support Vector Machines, Random Forests, Decision Trees, and Naïve Bayes. Their experimental results demonstrated the ability of their proposed combinations to improve predictive performance against all base classifiers. Their model was validated over five real-world credit scoring datasets.

Musyoka (6) compared data mining algorithms with the credit card approval dataset. This research focused on masked attributes and compared the Bayesian Network, Decision Tree, and J48 classifiers. Musyoka (6)'s results identified the Bayesian Network algorithm as being the most accurate, returning an accuracy of 86.21%.

Tanikella (7) examined the key features considered for issuing credit cards to customers. This work used machine learning to find that the attributes, prior default, years employed, credit score and debt were the most useful features.

Zhao (8) analyzed the prediction accuracy of multiple regression models and classifiers based on predetermined performance criterion. The experimental models used were Logistic Regression, Linear Support Vector Classification (Linear SVC), and the Naïve Bayes Classifier. In this study, linear SVC performed the best.

Though quite a few works, as presented above, have been done on different aspects of credit analysis using machine learning, none of the works have used the concept of synthetic feature generation in machine learning for credit analysis, which is the uniqueness and novelty of this article.

3. Datasets and processing

Four datasets were selected for this research: German Credit Risk (9), Credit Screening (10), Credit (11), and Bank Churners (12). All datasets contained attributes or features relating to credit cards or credit limits. In the tables describing the respective datasets, the attributes that appear in all four datasets are identified with four asterisks (****), the attributes that appear in three datasets are identified with three asterisks (***), and the attributes that appear in two datasets are identified with two asterisks (**). Preprocessing played a major role in this work, hence preprocessing is explained in detail in this section.

3.1. Preprocessing using feature selection

Feature selection is the process of identifying and selecting features or attributes within the dataset that will aid in improving the accuracy of the returned results. The selection process can be manual or automatic, but essentially the objective is the same – to achieve higher predictive accuracy. For this research, both manual and automatic feature selection was used in each dataset. Once the irrelevant or unusable attributes were removed, the datasets were imported into Weka, and Information Gain was run on each dataset using the Ranker search method. The output identified the amount of information gain for each attribute. Information gain is an entropy-based algorithm that determines the most relevant features necessary of the classification of a dataset.

3.2. Dataset 1: German credit risk

The German credit risk, obtained from Kaggle, was provided by Hofmann (9). This dataset consisted of 1,000 instances and ten attributes. Attribute descriptions and sample values are presented in **Table 1**.

3.2.1. Preprocessing the German credit risk dataset

3.2.1.1. Calculating information gain. For preprocessing, first the information gain was calculated using the original attributes. As shown in **Figure 1**, in the German Credit Risk dataset, the attribute with the highest information gain

TABLE 1 | Dataset 1: German credit risk (9).

No	Attribute	Description	Values
1.	Age****	Age of customer	Continuous
2.	Checking acct	Customer's checking account balance in Deutsch Marks	Little/moderate/rich
3.	Credit***	Customer's credit amount in Deutsch Marks	Continuous
4.	Duration	Length of customer's credit with bank	Continuous
5.	Gender****	Customer's gender	Male/female
6.	Housing	Customer's housing type	Own/rent/free
7.	Job	Customer's job type	0,1,2,3
8.	Purpose	Customer's reason for needing credit	Various
9.	Row number	Unique row number	Various
10.	Saving account	Customer's savings account level/balance	Little/moderate/rich/quite rich

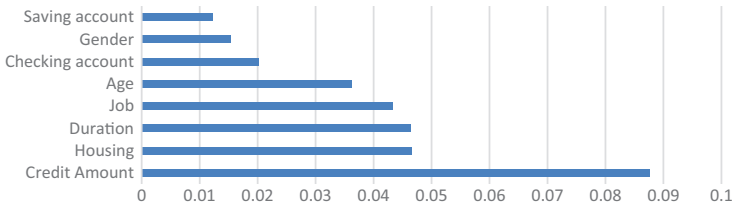

FIGURE 1 | Information gain for German credit risk dataset (dataset 1).

was credit amount, followed by housing and duration. The attributes that are not in **Figure 1** have information gain values very close to zero.

3.2.1.2. Removing attributes. The attribute, Purpose, was omitted. Based on its description and data it contained, it was not deemed relevant for this study.

3.2.1.3. Binning and synthetic feature generation. For the German Credit Risk dataset (9), synthetic feature (SF) generation was utilized for the attributes that had lower information gain: Age, Checking Account, Duration, and Saving Account.

The attribute, Age, was binned in two ways, as shown in **Table 2**. For regular binning, Age was grouped into four buckets, and for synthetic feature generation, the groups were based on the accepted classifications of the age generations (13). **Figures 2**, **3** show the distributions of each of the binning criteria.

TABLE 2 | Age: binning and SF generation.

Age (binned)	Age (SF)	
<25	Gen alpha	<18
25–35	Gen Z	18–22
36–45	Millennials	23–38
>45	Gen X	39–54
	Baby boomers	55–73
	Silent Generation	>73

The attribute, Job, was binned in two ways, as shown in **Table 3**.

The attributes, Checking and Savings, were binned as per **Table 4**.

The Duration attribute was grouped into four buckets and was narrowed down to three buckets for synthetic feature generation, as shown in **Table 5**.

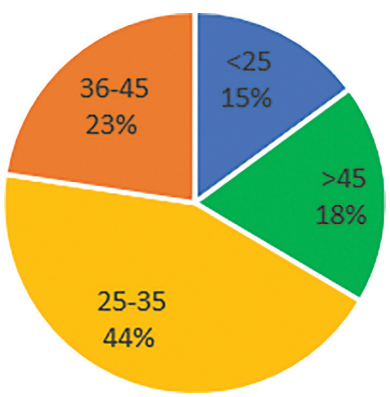

FIGURE 2 | Age (binned).

TABLE 3 | Job: binning and SF generation.

Job (binned)	Job (SF)
Unskilled	Unskilled
Skilled	Skilled
Highly skilled	

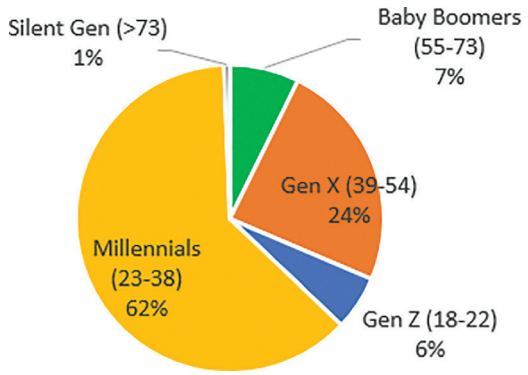

FIGURE 3 | Age (SF).

The Credit attribute was grouped into four buckets and was narrowed down to three buckets, for synthetic feature generation, as shown in Table 6.

3.3. Dataset 2: Credit screening

The Credit Screening dataset was obtained from UCI Machine Learning Repository and provided by Keogh et al. (10). The dataset contained 16 variables but the variables were masked, hence the converted attributes were used as per Rane (14)'s. This dataset consisted of 690 labeled instances. A description of the attributes is presented in Table 7.

3.3.1. Preprocessing the credit screening dataset

3.3.1.1. Calculating information gain. Information gain was calculated using the original set of attributes. As shown in Figure 4, in the Credit Screening dataset (10), the attribute with the highest information gain was Prior Default, and the attribute with the second highest information gain was credit score, with Employed following closely behind. The attributes that are not in Figure 4 have information gain values very close to zero.

3.3.1.2. Removing attributes. Citizen, education level, ethnicity, and zip code were removed.

- Citizen had three values: g, p, and s; g accounted for 90.5% of the applications so the assumption was made the most of the applicants were citizens and therefore the attribute was not used.
- The Education Level attribute had 14 unique values in alpha form, which were not easily interpretable, hence the attribute was removed (not used).
- The Zip Code attribute had values between 1 and 4 digits, hence was not used.
- Ethnicity had too many values, some of which were inconsistent; hence this attribute was removed (not used).

TABLE 4 | Checking and savings: binning and SF generation.

Checking (binned)	Savings (binned)	SF
Little	Little	<Moderate
Moderate	Moderate	Moderate
Rich	Rich	>Moderate
NA	Quite rich	
	NA	

TABLE 5 | Duration (months): binning and SF generation.

Duration (binned)	Duration (SF)
<12	<12
12–18	12–24
19–24	>24
>24	

TABLE 6 | Credit: binning and SF generation.

Credit (binned)	Credit (SF)
<1500	Low
1500–3000	Medium
3001–5000	High
>5000	

TABLE 7 | Dataset 2: credit screening (10).

No	Attribute	Original value	Converted value
1.	Age****	Continuous	
2.	Bank customer	g, p, gg	
3.	Citizen	g, p, s	
4.	Credit approved	+, −	Yes, no
5.	Credit score	Continuous	
6.	Debt	Continuous	
7.	Driver's license	t, f	True, false
8.	Education***	c, d, cc, I, j, k, m, r, q, w, x, e, aa, ff	
9.	Employed	t, f	True, false
10.	Ethnicity**	v, h, bb, j, n, z, dd, ff, o	
11.	Gender****	a, b	a = Male; b = Female
12.	Income***	Continuous	
13.	Married***	u, y, l	No, yes, unknown
14.	Prior default	t, f	True, false
15.	Years employed	Continuous	
16.	Zip code	Continuous	

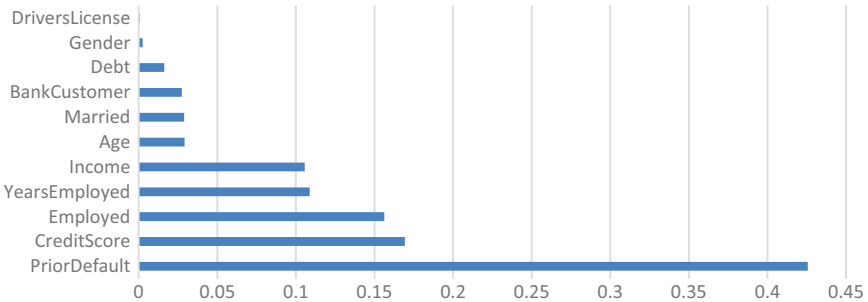

FIGURE 4 | Information gain on credit approved dataset (dataset 2).

3.3.1.3. Handling missing values. All missing values were labeled "unknown."

3.3.1.4. Binning and synthetic feature generation. Three of the attributes were binned: Age, Credit Score, and Income. Age was grouped as per Dataset 1 (**Table 2**), hence is not shown here. The binning and synthetic feature generation of Credit Score and Income are shown in **Tables 8**, **9**, respectively. Debt and years employed are binned in **Tables 10**, **11**, respectively.

3.4. Dataset 3: Credit

The third dataset, Credit, obtained from Kaggle.com, was provided by Iacob (11). This dataset contained 400 instances with 11 attributes, as shown in **Table 12**. This research focuses on the effects of the attributes on credit limit.

TABLE 8 | Credit score: binning and SF generation.

Credit score (binned)	Credit score (SF)
<1	<1
1	≥1
>1	

TABLE 9 | Income: binning and SF generation.

Income (binned)	Income (SF)
<1	<1
1–499	≥1
500–2000	
>2000	

TABLE 10 | Debt: binning.

Debt (binned)
<3
3–6
>6

3.4.1. Preprocessing the credit dataset

3.4.1.1. Calculating information gain. Information Gain was calculated on the original attributes. From **Figure 5**, it can be noted that the attribute with the highest information gain was monthly balance, with credit rating being the second highest. There are far less attributes in **Figure 5** than in **Table 12**. The attributes that are not in **Figure 5** have information gain values very close to zero.

3.4.1.2. Removing attributes. Student, ethnicity, income, ID, and number of cards were removed.

- The student attribute was removed because the other datasets did not contain a similar attribute and only 10% of the cardholders were students.
- Ethnicity attribute was also removed because it was not adequately identified in the other datasets.
- Income data did appear correct, hence was not used.
- ID attribute was removed.
- Number of cards was not used because the information gain was close to zero, and other datasets did not include this attribute.

3.4.1.3. Binning and synthetic feature generation. For the Credit Limit dataset, synthetic feature creation was applied on two attributes: Age and Credit Limit. Age was grouped in the same buckets as the German Credit (**Table 2**). Credit Limit was grouped once in numerical buckets and the second grouping (synthetic feature) was High, Medium, and Low, as shown in **Table 13**. Other attributes that were binned were monthly balance (**Table 14**), credit rating (**Table 15**), and education (**Table 16**).

TABLE 11 | Years employed: binning.

Years employed (binned)
<1
1–2
>2

TABLE 12 | Dataset 3: credit (11).

No	Attribute	Description	Values
1.	Age****	Age of cardholder	Continuous
2.	Balance**	Cardholder's average credit card balance in dollars	Continuous
3.	Credit limit***	Credit limit assigned by card company	Continuous
4.	Credit rating	Cardholder's credit rating	Continuous
5.	Education***	Number of educational years	5 through 20
6.	Ethnicity**	Cardholder's ethnicity	African American, Asian, Caucasian
7.	Gender****	Cardholder's gender	Female, male
8.	ID	Cardholder's identification	Continuous
9.	Income***	Cardholder's income in $10,000 increments	Continuous
10.	Married***	Cardholder's marital status	No, yes
11.	Number of cards	Cardholder's credit card count	1 through 9
12.	Student	Cardholder's current student status	No, yes

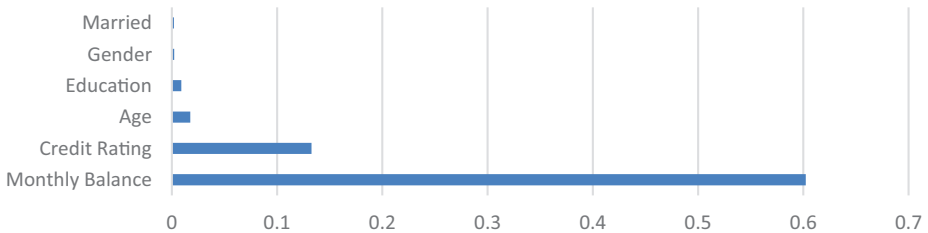

FIGURE 5 | Information gain on credit limit dataset (dataset 3).

3.5. Dataset 4: Bank churners dataset

The Bank Churners dataset, obtained from Kaggle.com, was provided by Goyal (12). This dataset contained 10,127 instances with 27 attributes, as shown in **Table 17**. The Bank Churners dataset focused on bank customers and the relationship between attrition and the other attributes in dataset.

TABLE 13 | Credit limit: binning and SF generation.

Credit limit (binned)	Credit limit (SF)
<1500	Low
1500–3000	Medium
3001–5000	High
>5000	

TABLE 14 | Balance: binning.

Balance (binned)
<350
350–700
>700

3.5.1. Preprocessing the bank churners dataset

3.5.1.1. Calculating information gain. Information gain was calculated using the original attributes. As shown in **Figure 6**, in the Bank Churners dataset, the attribute with the highest information gain was income, followed by gender and revolving balance. The attributes that are not in **Figure 6** have information gain values very close to zero.

TABLE 15 | Credit rating: binning.

Credit rating (binned)
<560
560–659
660–724
725–759
>759

TABLE 16 | Education: binning.

Education (binned)
<13
13–16
>16

TABLE 17 | Dataset 4: bank churners (12).

No	Attribute	Description	Values
1.	Age****	Age of customer	Continuous
2.	Attrition	Account closed or open	Existing/attritted
3.	Balance**	Revolving balance	Continuous
4.	Card category	Card category	Blue/gold/platinum/silver
5.	Client number	Unique identifier for each client	Continuous
6.	Contacts	Number of contacts in last 12 months	Continuous
7.	Credit limit***	Card credit limit	Continuous
8.	Dependents	Number of dependents	Continuous
9.	Education***	Education level	High school/college/graduate/doctorate/post-graduate/uneducated
10.	Gender****	Gender	Male/female
11.	Income***	Income	Categorized
12.	Married***	Marital status	Married/single/divorced
13.	Months w/Bank	Number of active months with bank	Continuous
14.	Months Inactive	Number of inactive months	Continuous
15.	Months total	Number of months with bank	Continuous
16.	Open to buy	Open to buy credit line	Continuous
17.	Total amt change	Changed in transaction amount (Q4 over Q1)	Continuous
18.	Total trans amount	Total transaction amount in last 12 months	Continuous
19.	Total trans count	Total transaction count in last 12 months	Continuous
20.	Diff trans cnt qtrly	Difference in transaction count (Q4 over Q1)	Continuous
21.	Utilization ratio	Average card utilization ratio	Continuous

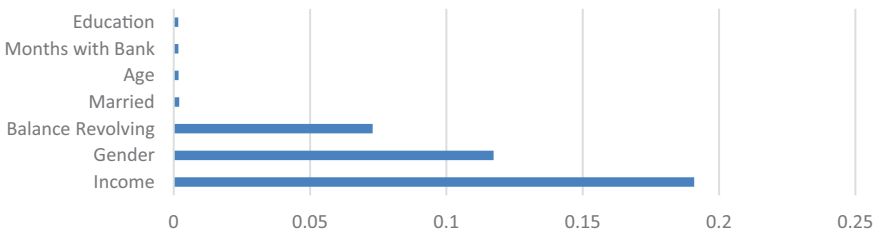

FIGURE 6 | Information gain on bank churners dataset (dataset 4).

3.5.1.2. Removing attributes. Attrition, Card Category, Client Number, Contacts, Dependents, Months Inactive, Months Total, Open to Buy, Total Amount Change, Total Transaction Amount, Total Transaction Count, Difference Transaction Count Quarterly, and Utilization Ratio Average were removed since they were not considered relevant for this analysis.

3.5.1.3. Binning and synthetic feature generation. For Dataset 4, Synthetic Feature Creation was done for Age (grouped as per the German Credit dataset, **Table 2**), Credit Limit (grouped as per the Credit dataset, **Table 13**), and Married. The Married attribute contained the following values: divorced, married, single, and unknown. Divorced and Single were grouped together into "no" (not married), leaving the grouped values for the Married attribute to be Yes, No, and Unknown, as shown in **Table 18**. Income was binned as per **Table 19**, and the synthetic feature for income was also categorized as shown in **Table 19**. Balance Revolving was binned as per **Table 20**, and months with bank were binned as per **Table 21**.

4. Classifiers used

For classification, three tree-based classifiers (Random Forest, Random Tree, and REPTree) and

TABLE 18 | Married: binning and SF generation.

Married (binned)	Married (SF)
Divorced	No
Married	Yes
Single	unknown
Unknown	

TABLE 19 | Income: binning and SF generation.

Income (original bin)	Income (SF)
<$40,000	<$60,000
$40,000–$60,000	$60,000–$120,000
$60,000–$80,000	$120,000+
$80,000–$120,000	Unknown
$120,000+	
Unknown	

TABLE 20 | Balance revolving: binning.

Balance revolving (binned)
<1000
1000–2000
>2000

TABLE 21 | Months with bank: binning.

Months with bank (binned)
<24
24–35
36–47
>47

one non-tree-based classifier (Naïve Bayes) were used.

4.1. Random forest

Random Forest is a widely used machine learning classifier that constructs multiple decision trees randomly, and the term "forest" stems from the imagery of the many trees being created. In Random Forest, each tree is independently produced without pruning, and the nodes are split based on the user's selection of available features (15). There are also works on how a user can prune Random Forest. Kulkarni and Sinha (16) showed a way of pruning by limiting the number of trees.

4.2. Random tree

The Random Tree classifier is similar to the Random Forest classifier, but it constructs only one decision tree and is based on a random set of attributes. The Random Tree classifier constructs a set of data to build the Random Tree, and every node is split from the best split among all variables (17). Essentially, the Random Tree is a simpler tree/forest classifier, but Random Forest tends to have better accuracy by decreasing the variance because it constructs multiple trees.

4.3. The REPTree classifier

Reduced error pruning tree (REPTree) builds the decision tree based on information gain (18). The tree that is built may be a decision/regression tree, but it is used for classification, and it will create multiple trees in different iterations (18). When the algorithm runs, it goes from each node starting at the bottom and works its way to the top, and at each node, it assesses if it should replace it with the most frequent class to improve the accuracy, and it prunes away items that would cause a reduction in accuracy (19). REPTree only sorts numeric attributes once.

4.4. The Naïve Bayes classifier

Naïve Bayes was chosen as an additional classifier because it is not a tree classifier. The Naïve Bayes classifier assumes that all variables are independent of the class attribute (20).

5. Results and discussion

The four different classifiers were run using Weka. For each dataset and each classifier, we looked at the following statistical measures: accuracy, true positive rate (TPR), false positive rate (FPR), precision, F-measure, and ROC area.

TABLE 22 | Naïve Bayes classification using the German credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
8 (4 OR; 4 Bin; 0 SF)	46.1%	46.1%	21.1%	43.2%	43.5%	71.2%	1, 2, 3, 4, 5, 6, 7, 9
8 (4 OR; 3 Bin; 1 SF)	72.8%	72.8%	39.9%	72.0%	72.3%	79.5%	1, 2, 3SF, 4, 5, 6, 7, 9
5 (1 OR; 3 Bin; 1 SF)	72.6%	72.6%	39.5%	72.0%	72.2%	79.1%	1, 3SF, 4, 6, 7
10 (4 OR; 2 Bin; 3 SF)	73.6%	73.6%	32.7%	74.5%	74.0%	79.5%	1, 1SF, 2, 3SF, 4, 4SF, 5, 6, 7, 9
11 (2 OR; 3 Bin; 6 SF)	72.3%	72.3%	36.0%	72.7%	72.5%	78.6%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 7, 7SF, 9, 9SF
12 (3 OR; 3 Bin; 6 SF)	72.5%	72.5%	35.3%	73.1%	72.7%	78.9%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 5, 7, 7SF, 9, 9SF
13 (4 OR; 3 Bin; 6 SF)	72.5%	72.5%	35.0%	73.2%	72.8%	78.9%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 5, 6, 7, 7SF, 9, 9SF

TABLE 23 | Random forest classification using the German credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
8 (4 OR; 4 Bin; 0 SF)	38.0%	38.0%	23.0%	38.0%	37.4%	65.8%	1,2,3,4,5,6,7,9
8 (4 OR; 3 Bin; 1 SF)	71.6%	71.6%	40.5%	71.1%	71.3%	75.5%	1, 2, 3SF, 4, 5, 6, 7, 9
5 (1 OR; 3 Bin; 1 SF)	73.2%	73.2%	40.0%	72.3%	72.6%	76.7%	1, 3SF, 4, 6, 7
10 (4 OR; 2 Bin; 3 SF)	71.5%	71.5%	41.1%	70.8%	71.1%	75.3%	1, 1SF, 2, 3SF, 4, 4SF, 5, 6, 7, 9
11 (2 OR; 3 Bin; 6 SF)	71.5%	71.5%	43.1%	70.4%	70.7%	74.9%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 7, 7SF, 9, 9SF
12 (3 OR; 3 Bin; 6 SF)	70.7%	70.7%	42.9%	69.8%	70.2%	74.3%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 5, 7, 7SF, 9, 9SF
13 (4 OR; 3 Bin; 6 SF)	71.5%	71.5%	41.4%	70.7%	71.1%	75.0%	1, 1SF, 2, 2SF, 3SF,4, 4SF, 5, 6, 7, 7SF, 9, 9SF

TABLE 24 | Random tree classification using the German credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
8 (4 OR; 4 Bin; 0 SF)	37.6%	37.6%	23.4%	35.1%	35.8%	59.9%	1,2,3,4,5,6,7,9
8 (4 OR; 3 Bin; 1 SF)	68.7%	68.7%	40.3%	69.4%	69.0%	67.5%	1, 2, 3SF, 4, 5, 6, 7, 9
5 (1 OR; 3 Bin; 1 SF)	72.3%	72.3%	38.7%	72.0%	72.1%	75.4%	1, 3SF, 4, 6, 7
10 (4 OR; 2 Bin; 3 SF)	69.9%	69.9%	39.2%	70.4%	70.1%	67.9%	1, 1SF, 2, 3SF, 4, 4SF, 5, 6, 7, 9
11 (2 OR; 3 Bin; 6 SF)	69.2%	69.2%	41.4%	69.3%	69.3%	70.0%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 7, 7SF,9, 9SF
12 (3 OR; 3 Bin; 6 SF)	68.0%	68.0%	41.7%	68.5%	68.2%	66.4%	1, 1SF, 2, 2SF, 3SF, 4, 4SF, 5, 7, 7SF, 9, 9SF
13 (4 OR; 3 Bin; 6 SF)	69.6%	69.6%	39.5%	70.1%	69.8%	67.3%	1, 1SF, 2, 2SF, 3SF,4, 4SF, 5, 6, 7, 7SF, 9, 9SF

TABLE 25 | REPTree classification using the German credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
8 (4 OR; 4 Bin; 0 SF)	45.7%	45.7%	21.6%	41.8%	40.3%	68.0%	1,2,3,4,5,6,7,9
8 (4 OR; 3 Bin; 1 SF)	72.7%	72.7%	40.0%	71.9%	72.2%	76.4%	1,2,3SF,4,5,6,7,9
5 (1 OR; 3 Bin; 1 SF)	72.5%	72.5%	44.5%	70.9%	71.2%	76.7%	1,3SF,4,6,7
10 (4 OR; 2 Bin; 3 SF)	71.8%	71.8%	42.6%	70.7%	71.1%	75.6%	1,1SF,2,3SF,4,4SF,5,6,7,9
11 (2 OR; 3 Bin; 6 SF)	72.2%	72.2%	43.0%	70.9%	71.3%	76.5%	1,1SF,2,2SF,3SF,4,4SF,7,7SF,9, 9SF
12 (3 OR; 3 Bin; 6 SF)	71.8%	71.8%	42.9%	70.6%	71.0%	75.5%	1,1SF,2,2SF,3SF,4,4SF,5,7,7SF, 9, 9SF
13 (4 OR; 3 Bin; 6 SF)	71.8%	71.8%	42.6%	70.7%	71.1%	75.6%	1,1SF,2,2SF,3SF,4,4SF,5,6,7,7SF, 9,9SF

TABLE 26 | Classifier accuracy comparison on the German credit dataset.

	Classification accuracy				
No. of attributes	Naïve Bayes	Random forest	Random tree	REPTree	Attributes used
8 (4 OR; 4 Bin; 0 SF)	46.1%	38.0%	37.6%	45.7%	1,2,3,4,5,6,7,9
8 (4 OR; 3 Bin; 1 SF)	72.8%	71.6%	68.7%	**72.7%**	1,2,3SF,4,5,6,7,9
5 (1 OR; 3 Bin; 1 SF)	72.6%	**73.2%**	**72.3%**	72.5%	1,3SF,4,6,7
10 (4 OR; 2 Bin; 3 SF)	**73.6%**	71.5%	69.9%	71.8%	1,1SF,2,3SF,4,4SF,5,6,7,9
11 (2 OR; 3 Bin; 6 SF)	72.3%	71.5%	69.2%	72.2%	1,1SF,2,2SF,3SF,4,4SF,7,7SF,9, 9SF
12 (3 OR; 3 Bin; 6 SF)	72.5%	70.7%	68.0%	71.8%	1,1SF,2,2SF,3SF,4,4SF,5,7,7SF, 9, 9SF
13 (4 OR; 3 Bin; 6 SF)	72.5%	71.5%	69.6%	71.8%	1,1SF,2,2SF,3SF,4,4SF,5,6,7,7SF, 9,9SF

Accuracy is the ratio of a model's correct data (TP + TN) to the total data, calculated by the following equation:

$$(TP + TN)/(TP + TN + FP + FN).$$

TPR, also called sensitivity or recall, measures the proportion of actual attacks that were identified as attacks, given by the following equation:

$$TP/(TP + FN).$$

TABLE 27 | German credit dataset – improvement in accuracy with synthetic features.

Classifier	Synthetic feature creation	Accuracy	FP rate	Accuracy improvement
Naïve Bayes	Yes	73.6%	32.7%	27.5%
	No	46.1%	21.1%	
Random forest	Yes	73.2%	40.0%	35.2%
	No	38.0%	23.0%	
Random tree	Yes	72.3%	38.7%	34.7%
	No	37.6%	24.1%	
REPTree	Yes	72.7%	40.0%	27%
	No	45.7%	21.8%	

FPR is where a non-attack is identified as an attack, given by the following equation:

$$FP/(FP + TN).$$

Precision measures the proportion of positive identification of attacks that were actually attacks, given by the following equation:

$$TP/(TP + FP).$$

F-measure is the harmonic means on precision and recall, calculated by the following equation:

$$2^*((Precision^*Recall)/(Precision + Recall))$$

ROC plots the relational of the TPR vs. FPR.
Where:

- True Positive (TP) is instances that were identified correctly as positives.
- True Negative (TN) is instances that were identified correctly as negatives.
- False Positive (FP) is instances that were identified incorrectly as positives.
- False Negative (FN) is instances that were identified incorrectly as negatives.

For each of the classification runs, various combinations of original attributes (OR), original binned attributes (Bin), and synthetic features (SF) were used. The original attributes and original binned attributes were selected based on information gain that was performed on each respective dataset.

To select the best results, the runs with the minimal set of attributes with the highest statistical measures were selected.

5.1. Classification results for the German credit dataset

For the German credit dataset (9), credit amount, a continuous attribute, was used as the class variable for classification. Tables 22–25 present the statistical results of the classifications.

5.1.1. Naïve Bayes results

Results of the Naïve Bayes classification, presented in Table 22, show that 10 attributes with three synthetic features had the best results, with a classification accuracy of 73.6%. In this run, three synthetic features were used, for features, age, credit, and duration. The other statistical measures for this run were also high. Without the use of any synthetic features, for this dataset, the classification results were really poor (accuracy 46.1%).

5.1.2. Random forest results

Results of the Random Forest classification, presented in Table 23, show that five attributes with one synthetic feature had the best results in terms of classification accuracy (73.2%). In this run, only the synthetic feature for credit was used. The other statistical measures for this run were also high. Without the use of any synthetic features, for this dataset, the classification results were really poor (accuracy 38%).

TABLE 28 | Naïve Bayes classification using the credit screening dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
10 (5 OR; 5 BIN)	84.6%	84.6%	16.4%	84.7%	84.6%	91.3%	1,2,3,5,6,9,12,13,14,15
12 (7 OR; 5 BIN)	84.8%	84.8%	16.2%	84.8%	84.7%	91.2%	1,2,3,5,6,7,9,11,12,13,14,15
13 (5 OR; 5 BIN; 3 SF)	81.6%	81.6%	19.7%	81.6%	81.5%	89.4%	1,1SF,2,3,5,5SF,6,9,12,12SF,13,14,15

TABLE 29 | Random forest classification using the credit screening dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
10 (5 OR; 5 BIN)	83.3%	83.3%	17.0%	83.4%	83.3%	90.9%	1,2,3,5,6,9,12,13,14,15
12 (7 OR; 5 BIN)	84.5%	84.5%	15.9%	84.5%	84.5%	91.0%	1,2,3,5,6,7,9,11,12,13,14,15
13 (5 OR; 5 BIN; 3 SF)	85.4%	85.4%	15.0%	85.4%	85.4%	91.2%	1,1SF,2,3,5,5SF,6,9,12,12SF,13,14,15

TABLE 30 | Random tree classification using the credit screening dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
10 (5 OR; 5 BIN)	82.2%	82.2%	17.6%	82.4%	82.2%	85.5%	1,2,3,5,6,9,12,13,14,15
12 (7 OR; 5 BIN)	82.2%	82.2%	18.2%	82.2%	82.2%	83.5%	1,2,3,5,6,7,9,11,12,13,14,15
13 (5 OR; 5 BIN; 3 SF)	83.8%	83.8%	16.1%	83.9%	83.8%	85.6%	1,1SF,2,3,5,5SF,6,9,12,12SF, 13,14,15

TABLE 31 | REPTree classification using the credit screening dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes used
10 (5 OR; 5 BIN)	84.8%	84.8%	14.7%	85.1%	84.8%	89.4%	1,2,3,5,6,9,12,13,14,15
12 (7 OR; 5 BIN)	84.3%	84.3%	15.2%	84.7%	84.4%	89.5%	1,2,3,5,6,7,9,11,12,13,14,15
13 (5 OR; 5 BIN; 3 SF)	84.6%	84.6%	15.3%	84.8%	84.7%	88.7%	1,1SF,2,3,5,5SF,6,9,12,12SF,13,14,15

TABLE 32 | Classifier accuracy comparison on credit screening dataset.

	Accuracy				
No. of attributes	Naïve Bayes	Random forest	Random tree	REPTree	Attributes used
10 (5 OR; 5 BIN)	84.6%	83.3%	82.2%	**84.8%**	1,2,3,5,6,9,12,13,14,15
12 (7 OR; 5 BIN)	**84.8%**	84.5%	82.2%	84.3%	1,2,3,5,6,7,9,11,12,13,14,15
13 (5 OR; 5 BIN; 3 SF)	81.6%	**85.4%**	**83.8%**	84.6%	1,1SF,2,3,5,5SF,6,9,12,12SF,13,14,15

5.1.3. Random tree results

Results of the Random Tree classification, presented in Table 24, show that five attributes with one synthetic feature had the best results in terms of classification accuracy (72.3%). In this run, the only synthetic feature used was for credit. The other statistical measures for this run were also high. Again, without the use of any synthetic features, for this dataset, the classification results were really poor (accuracy 37.6%).

5.1.4. REPTree results

Results of the REPTree classification, presented in Table 25, show that eight attributes with one synthetic feature had slightly higher classification accuracy (72.7%) than the other runs. In this run, the only synthetic feature used was credit. Without the use of any synthetic features, for this dataset, the classification results were really poor (accuracy 45.7%).

TABLE 33 | Credit screening dataset – improvement in accuracy with synthetic features.

Classifier	Synthetic feature creation	Accuracy	FP rate	Accuracy improvement
Naïve Bayes	Yes	81.6%	16.0%	−3.2%
	No	84.8%	16.2%	
Random forest	Yes	85.4%	15.1%	0.9%
	No	84.5%	15.9%	
Random tree	Yes	83.8%	18.8%	1.6%
	No	82.2%	18.2%	
REPTree	Yes	84.6%	15.2%	−0.2%
	No	84.8%	15.2%	

5.1.5. Overall classifier comparison for the German credit dataset

From Tables 22–25, it can be noted that even using one synthetic attribute greatly improved the classification accuracy and other statistical measures.

A comparison of the classification accuracy of the all the classifiers, on the German Credit dataset (Table 26), show that two out of the four classifiers performed well with five attributes and only one synthetic feature. Naïve Bayes performed the best with 10 attributes and three synthetic features, and REPTree performed the best with eight attributes and one synthetic feature. The highest classification accuracy was achieved with the Naïve Bayes classifier, and the most improved accuracy was achieved using the Random Forest classifier (35.2% improvement, as shown in Table 27).

5.2. Classification results for the credit screening dataset

For the Credit Screening dataset (10), the attribute approved, a binary attribute, was used as the class variable for

TABLE 34 | Naïve Bayes classification using the credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (3 OR; 5 BIN)	59.0%	59.0%	18.2%	59.2%	58.4%	81.0%	1,2,3,4,5,7,10,12
5 (0 OR; 4 BIN; 1 SF)	77.0%	77.0%	22.8%	81.8%	78.3%	84.5%	1,2,3SF,4,5
6 (0 OR; 4 BIN; 2 SF)	76.3%	76.3%	24.4%	80.9%	77.6%	82.9%	1,1SF,2,3SF,4,5
7 (2 OR; 4 BIN; 1 SF)	77.5%	77.5%	24.7%	81.3%	78.6%	84.6%	1,2,3SF,4,5,7,10

TABLE 35 | Random forest classification using the credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (3 OR; 5 BIN)	54.0%	54.0%	20.0%	54.8%	54.2%	79.1%	1,2,3,4,5,7,10,12
5 (0 OR; 4 BIN; 1 SF)	75.8%	75.8%	35.8%	77.3%	76.4%	82.8%	1,2,3SF,4,5
6 (0 OR; 4 BIN; 2 SF)	77.5%	77.5%	40.1%	77.2%	77.3%	84.6%	1,1SF,2,3SF,4,5
7 (2 OR; 4 BIN; 1 SF)	76.3%	76.3%	39.1%	76.7%	76.4%	84.4%	1,2,3SF,4,5,7,10

TABLE 36 | Random tree classification using the credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (3 OR; 5 BIN)	54.3%	54.3%	19.9%	56.8%	54.9%	73.4%	1,2,3,4,5,7,10,12
5 (0 OR; 4 BIN; 1 SF)	75.3%	75.3%	41.6%	75.5%	75.4%	82.5%	1,2,3SF,4,5
6 (0 OR; 4 BIN; 2 SF)	76.8%	76.8%	43.9%	75.9%	76.3%	83.4%	1,1SF,2,3SF,4,5
7 (2 OR; 4 BIN; 1 SF)	75.5%	75.5%	45.7%	74.7%	75.0%	81.3%	1,2,3SF,4,5,7,10

TABLE 37 | REPTree classification using the credit dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (3 OR; 5 BIN)	59.5%	59.5%	16.9%	–	–	79.8%	1,2,3,4,5,7,10,12
5 (0 OR; 4 BIN; 1 SF)	79.3%	79.3%	15.8%	85.0%	80.5%	84.1%	1,2,3SF,4,5
6 (0 OR; 4 BIN; 2 SF)	78.3%	78.3%	16.1%	84.6%	79.6%	83.5%	1,1SF,2,3SF,4,5
7 (2 OR; 4 BIN; 1 SF)	76.3%	76.3%	31.4%	78.7%	77.1%	83.5%	1,2,3SF,4,5,7,10

TABLE 38 | Classifier accuracy comparison on credit dataset.

	Classification accuracy				
No. of attributes	Naïve Bayes	Random forest	Random tree	REPTree	Attributes used
8 (3 OR; 5 BIN)	59.0%	54.0%	54.3%	59.5%	1,2,3,4,5,7,10,12
7 (2 OR; 4 BIN; 1 SF)	75.8%	**78.0%**	**78.5%**	**78.8%**	1,2,3SF,4,5,7,10
5 (0 OR; 4 BIN; 1 SF)	76.3%	77.5%	76.8%	78.3%	1,2,3SF,4,5
6 (0 OR; 4 BIN; 2 SF)	**77.5%**	76.3%	75.5%	76.3%	1,1SF,2,3SF,4,5

classification. Tables 28–31 present the statistical results of the classifications. This dataset performed pretty well without the synthetic features too. Using three synthetic features for age, credit score, and income only very slightly improved the classification accuracy for the Random Forest and Random Tree algorithms.

5.2.1. Overall classifier comparison for the credit screening dataset

A comparison of the classification accuracy of the all the classifiers, on the Credit Screening dataset, from Table 32, shows that Random Forest and Random Tree had the highest accuracy using three synthetic features, but Naïve

TABLE 39 | Credit dataset – improvement in accuracy with synthetic features.

Classifier	Synthetic feature creation	Accuracy	FP rate	Accuracy improvement
Naïve Bayes	Yes	77.5%	26.1%	18.5%
	No	59.0%	18.2%	
Random forest	Yes	78.0%	42.8%	24%
	No	54.0%	20.0%	
Random tree	Yes	78.5%	44.7%	24.2%
	No	54.3%	19.9%	
REPTree	Yes	78.8%	15.8%	19.3%
	No	59.5%	17.6%	

Bayes and REPTree did not perform better with synthetic features. An analysis of the accuracy improvement on this dataset, as shown in **Table 33**, shows very little improvement after adding synthetic features. In fact, there was a negative improvement with Naïve Bayes and REPTree.

5.3. Classification results for the credit dataset

For the Credit dataset (11), credit limit was used as the class variable for classification. **Tables 34–37** present the statistical results of the classifications.

5.3.1. Naïve Bayes results

Results of the Naïve Bayes classification, presented in **Table 34**, show that in terms of classification accuracy, seven attributes with one synthetic feature, credit limit, had the best results, with a classification accuracy of 77.5%. Other statistical measures were also high for this run.

5.3.2. Random forest results

Results of the Random Forest classification, presented in **Table 35**, show that six attributes with two synthetic features, age and credit limit, had the best results in terms of classification accuracy (77.5%).

5.3.3. Random tree results

Results of the Random Tree, presented in **Table 36**, show that six attributes with two synthetic features, age and credit limit, had the best results in terms of classification accuracy (76.8%).

5.3.4. REPTree results

Results of the REPTree classification, presented in **Table 37**, show that five attributes with one synthetic feature, credit limit, had the best results in terms of classification accuracy (79.3%).

5.3.5. Overall classifier comparison for the credit dataset

For the Credit dataset, for all classifiers, there was a significant increase in classification accuracy and other statistical measures after the synthetic features were added, as shown in **Table 38**. Comparing the classifiers, REPTree performed the best at 78.8% classification accuracy. Three of the four classifiers performed the best with seven attributes and one synthetic feature. Only Naïve Bayes performed the best with six attributes and two synthetic features. Other statistical measures were also higher for this set of runs.

From **Table 39**, it can be observed that Random Tree had the highest improvement in accuracy (24.2%), closely followed by Random Forest at 24%. The other two classifiers, Naïve Bayes and REPTree, also had a significant improvement with the addition of synthetic attributes.

5.4. Classification results for the bank churners dataset

For the Bank Churners dataset (12), credit limit was used as the class variable for classification. **Tables 40–43** present the statistical results of the classifications.

Results of the Naïve Bayes classification, presented in **Table 40**, show that in terms of classification accuracy, four attributes with one synthetic feature, credit limit, had the best results, with a classification accuracy of 71.1%. Results of the Random Forest, Random Tree, and REPTree, presented in **Tables 41–43**, respectively, also show that four attributes with one synthetic feature, credit limit, had the best results in terms of classification accuracy (72.7, 72.7, and 72.5%, respectively).

TABLE 40 | Naïve Bayes classification using the bank churners dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (4 OR; 4 BIN)	56.9%	56.9%	25.7%	52.3%	51.1%	73.8%	1,3,7,9,10,11,12,13
4 (2 OR; 2 BIN)	57.2%	57.2%	26.3%	–	–	73.4%	3,7,10,11
4 (2 OR; 1 BIN; 1 SF)	71.1%	71.1%	30.7%	72.80%	71.60%	75.5%	3,7SF,10,11

TABLE 41 | Random forest classification using the bank churners dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (4 OR; 4 BIN)	52.0%	52.0%	26.5%	48.3%	49.5%	68.2%	1,3,7,9,10,11,12,13
4 (2 OR; 2 BIN)	57.7%	57.7%	25.5%	–	–	73.8%	3,7,10,11
4 (2 OR; 1 BIN; 1 SF)	72.7%	72.7%	35.2%	72.3%	72.4%	75.9%	3,7SF,10,11

TABLE 42 | Random tree classification using the bank churners dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (4 OR; 4 BIN)	50.0%	50.0%	26.3%	47.8%	48.7%	65.3%	1,3,7,9,10,11,12,13
4 (2 OR; 2 BIN)	57.7%	57.7%	25.5%	–	–	73.8%	3,7,10,11
4 (2 OR; 1 BIN; 1 SF)	72.7%	72.7%	35.2%	72.30%	72.40%	75.8%	3,7SF,10,11

TABLE 43 | REPTree classification using the bank churners dataset.

No. of attributes used	Accur.	TPR	FPR	Precision	F-measure	ROC	Attributes
8 (4 OR; 4 BIN)	56.0%	56.0%	25.9%	51.2%	52.2%	72.1%	1,3,7,9,10,11,12,13
4 (2 OR; 2 BIN)	57.8%	57.8%	25.9%	–	–	73.5%	3,7,10,11
4 (2 OR; 1 BIN; 1 SF)	72.5%	72.5%	36.8%	71.90%	72.10%	75.2%	3,7SF,10,11

TABLE 44 | Classifier accuracy comparisons on bank churners dataset.

	Classification Accuracy				
No. of attributes	Naïve Bayes	Random forest	Random tree	REPTree	Attributes used
4 (2 OR; 2 BIN)	57.2%	57.7%	57.7%	57.8%	3,7,10,11
8 (4 OR; 4 BIN)	56.9%	52.0%	50.0%	56.0%	1,3,7,9,10,11,12,13
4 (2 OR; 1 BIN; 1SF)	71.1%	72.7%	72.7%	72.5%	3,7SF,10,11

5.4.1. Overall classifier comparison for the bank churners dataset

For this set of classifiers, adding one synthetic attribute improved the classification accuracy significantly and all four classifiers performed the best at four attributes with one synthetic attribute, as shown in Table 44. From Table 45, it can be noted that Random Tree had the highest improvement in accuracy at 22.7%, followed by Random Forest at 20.7%. The other two algorithms also had significant improvement in accuracy with just one synthetic feature.

TABLE 45 | Bank churners–improvement in accuracy with synthetic features.

Classifier	Synthetic feature creation	Accuracy	FP rate	Accuracy improvement
Naïve Bayes	Yes	71.2%	30.9%	14.3%
	No	56.9%	25.7%	
Random forest	Yes	72.7%	35.2%	20.7%
	No	52.0%	26.5%	
Random tree	Yes	72.7%	35.2%	22.7%
	No	50.0%	26.3%	
REPTree	Yes	72.5%	36.8%	16.5%
	No	56.0%	25.9%	

6. Conclusion

Three of the four datasets used in this research showed an improvement in accuracy and other statistical measures using synthetic attributes. Overall, the tree-based classifiers, Random Forest, Random Tree, and REPTree, appeared to have better performances as well as better performance improvements than the non-tree-based classifier, Naïve Bayes.

Author contributions

SB conceptualized the article, responsible for guiding the research, and directing the formulation of the article. JW

also helped to conceptualize the article, did most of the pre-processing and processing of the data, and wrote the initial draft of the article. Both authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wongchinsri P, Kuratach W. A survey - data mining frameworks in credit card processing. *Proceedings of the 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. Chiang Mai: IEEE (2016). p. 1–6.
2. Zeng G. A necessary condition for a good binning algorithm in credit scoring. *Appl Math Sci.* (2014) 8:3229–42.
3. Danenas P, Garsva G. Selection of support vector machines-based classifiers for credit risk domain. *Expert Syst Appl.* (2015) 42:3194–204.
4. Lessmann S, Baesens B, Seow HV, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *Eur J Operat Res.* (2015) 247:124–36.
5. Ala'raj M, Abbod MF. Classifiers consensus system approach for credit scoring. *Knowl Based Syst.* (2016) 104:89–105.
6. Musyoka WM. Comparison of data mining algorithms in credit card approval. *Int J Comput Inform Technol.* (2018) 7:2.
7. Tanikella U. *Credit Card Approval Verification Model*. PhD thesis. San Marcos, CA: California State University San Marcos (2020).
8. Zhao Y. Credit card approval predictions using logistic regression, linear svm and naïve bayes classifier. *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*. Guilin: IEEE (2022). p. 207–11. doi: 10.1109/MLKE55170.2022.00047
9. Hofmann H. *German Credit Risk UCI Machine Learning*. San Francisco, CA: Kaggle (2016).
10. Keogh E, Blake C, Merz CJ. *UCI Repository of Machine Learning Databases*. (1998). Available online at: https://archive.ics.uci.edu/ml/datasets/credit+approval
11. Iacob S. *Predicting Credit Card Balance Using Regression*. San Francisco, CA: Kaggle (2020).
12. Goyal S. *Credit Card Customers Predict Churning Customers*. San Francisco, CA: Kaggle (2021).
13. Ricaldi LC. *Three Essays on Consumer Credit Card Behavior*. PhD thesis. Lubbock, TX: Texas Tech University (2015).
14. Rane K. *Credit Card Approval Analysis*. (2018). Available online at: https://nycdatascience.com/blog/student-works/credit-card-approval-analysis/
15. Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. *ISPRS J Photogr Remote Sens.* (2016) 114:24–31. doi: 10.1016/j.isprsjprs.2016.01.011
16. Kulkarni VY, Sinha PK. Pruning of random forest classifiers: a survey and future directions. *2012 International Conference on Data Science & Engineering (ICDSE)*. Cochin: IEEE (2012). p. 64–8. doi: 10.1109/ICDSE.2012.6282329
17. Mishra AK, Ratha BK. Study of random tree and random forest data mining algorithms for microarray data analysis. *Int J Adv Electr Comput Eng.* (2016) 3.
18. Kalmegh SR. Analysis of WEKA data mining algorithm reptree, simple cart and randomtree for classification of Indian news. *Int J Innov Sci Eng Technol.* (2015) 2.
19. Rokach L, Maimon O. *Data mining with decision trees - theory and applications*. 2nd ed. Singapore: World Scientific Publishing (2015).
20. Saritas MM, Yaşar AB. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *Int J Intell Syst Appl Eng.* (2019) 7:88–91.
21. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Patt Recogn.* (1997) 30:1145–59.
22. Breiman L. Random Forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
23. Frank E, Trigg LE, Holmes G, Witten IH. Naive bayes for regression (technical note). *Mach Learn.* (2000) 41:5–25. doi: 10.1023/A:1007670802811
24. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. Burlington, MA: Elsevier/Morgan Kaufmann (2012).
25. Seker S. *AutoML: End-to-end Introduction From Optiwisdom*. (2019). Available online at: https://towardsdatascience.com/automl-end-to-end-introduction-from-optiwisdom-c17fe03a017f
26. Taub J, Elliot M. The synthetic data challenge. *Conference of European Statisticians Joint UNECE/eurostat Work Session on Statistical Data Confidentiality*. Hague: UNECE (2019).
27. WEKA. *Weka 3: Machine Learning Software in Java, Weka 3 - Data Mining With Open Source Machine Learning Software in Java*. (2022). Available online at: https://www.cs.waikato.ac.nz/ml/weka/ (accessed on August 22, 2022).
28. Zhao H, Liu H, Fu Y. Incomplete multi-modal visual data grouping. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. Palo Alto, CA: AAAI Press (2016). p. 2392–8.