Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers

Ajay
B.E Computer Science and Engineering.
S.S.N College of Engineering
Kalavakam,Chennai,Tamil Nadu,India.

Ajay Venkatesh
B.E Computer Science and Engineering.
S.S.N College of Engineering
Kalavakam,Chennai,Tamil Nadu,India.

Shomona Gracia Jacob,
PhD
Associate Professor ME.,Ph.D.
S.S.N College of Engineering
Kalavakam,Chennai,Tamil Nadu,India.

ABSTRACT
Data mining is an emerging area of research that aims at extracting meaningful patterns from available data. This paper highlights the significance of classification in predicting new trends from voluminous data. Performance analysis of various data mining algorithm viz. BayesNet, Meta-Stacking, Naïve Bayes, Random Forest, SMO and ZeroR in predicting credit-card defaulters is discussed in this paper. Dataset from the UCI machine learning repository comprising of 25 attributes and 30000 instances have been employed to analyze the performance of algorithms. Moreover, the effect of feature selection has also been identified with respect to each classification algorithm. It has been concluded from the experimental results that both Correlation Feature Subset and Information Gain feature selection methods yield the most useful features for prediction and the accuracy of Random Forest Ensemble method is highest in predicting credit card defaulters.

General Terms
Data Mining, Credit card defaulters dataset.

Keywords
Credit card defaulter, datamining, patterns, Knowledge patterns.

1. INTRODUCTION
Data mining [1] includes the various techniques involved in exploring available data and recapitulating it into potentially useful information. Data Mining [2], in the field of computer science aims at detecting patterns and relationships among voluminous data in huge relational databases. These patterns then in turn act as input data for many machine learning algorithms. Data mining, classifies, groups, separates the given data. This can be used for predictive analysis with a decision support system. There are numerous data mining algorithms available. These have been wisely used for feature selection, classification, rule framing and clustering.

Use of datamining in banking sector has been on constant rise. Machine learning involves the usage of many classification algorithms, which is used to separate the data into many suggested number of categories.

With constant need for computer usage in banking sector, the increasing number of bank accounts created and credit outflows, computer can play the role of assistive banker in ensuring default less credits and safe banking. Among this credit default has been a major challenge for bankers as it endangers their system and pose chances of losses that might be hard to revert. Credit card defaulters are on the rise too.

Predicting the nature of a customer of whether he might be a defaulter or not is a complex function. Though the law has stringent measures against credit card defaulting, it is still prevalent in most parts. In India, credit card defaulters are charged under both civil and criminal cases.

In this research work, the performance measures of classification algorithm are compared on the Credit card default dataset published in UCI machine learning repository [3]. The effect of feature selection algorithms in analyzed.

1.1 Organization Of Paper
This paper is organized in this manner: Section 2 refers to the previous works on this dataset and credit default in general. Section 3 focuses on the dataset design and system description. Section 4 explains about the various feature selection algorithms used and their effects. Section 5 briefs about various classifying algorithms used.

2. RELATED WORK
There has been some intense research work on credit card default dataset. Abbas Keramati et.al [4] has done a literature survey on work that has been done on similar dataset. It does not extensively analyze any classification algorithm or effects of feature selection. Adela Ioana TUDOR et.al [5] dealt with clustering analysis on a similar dataset. In the above work, data is clustered relating to similarity among different attributes present. It does not evaluate or execute any classification algorithm or feature selection. The study of Simona Vasilica Oprea et.al [6] involves evaluation of few classification algorithms. They do not depict the change in efficiency of these algorithms with respect to feature selection.

3. DATASET AND SYSTEM DESIGN
Feature selection is the process of identifying the best subset of features that have the ability to categorize the given data under appropriate categories [16]Classification [2] is a phase in data mining that identifies items in a collection and places them under target categories. The training phase in classification aims at detecting relationships or associations between attributes and class values.

3.1 System Description
The diagrammatic representation of the system is shown in figure 1.

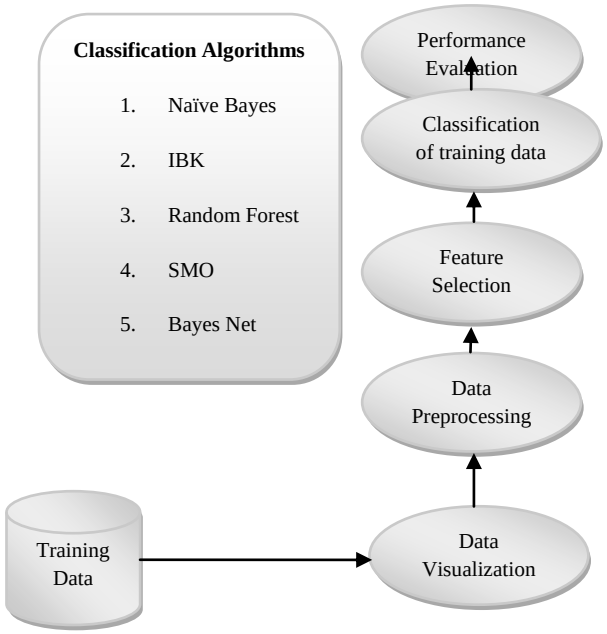

Figure 1.

3.1.1 Training data

The dataset is obtained from UCI Machine Learning Repository credit card defaulter [3]. It is a newly published dataset (obtained in 2015). The attribute details in the dataset are given in Table 1.

Table 1. Description of the attributes in the dataset

No	Attribute Name	Description
1	ID	User ID
2	Limit_Bal	Amount of the given credit(NT dollar)
3	Sex	Gender(1=male,2=female)
4	Education	Education (1=graduate school,2=University,3=others)
5	Marriage	Marital status(1=married,2=unmarried)
6	Age	Age(year)
7	Pay_0	the repayment status in September, 2005
8	Pay_2	the repayment status in August, 2005
9	Pay_3	the repayment status in July, 2005
10	Pay_4	the repayment status in June, 2005
11	Pay_5	the repayment status in May, 2005
12	Pay_6	the repayment status in April, 2005
13	Bill_Amt1	amount of bill statement in September,2005
14	Bill_Amt2	amount of bill statement in August,2005
15	Bill_Amt3	amount of bill statement in July,2005
16	Bill_Amt4	amount of bill statement in June,2005
17	Bill_Amt5	amount of bill statement in May,2005
18	Bill_Amt6	amount of bill statement in April,2005
19	Pay_Amt1	amount paid in September,2005
20	Pay_Amt2	amount paid in September,2005
21	Pay_Amt3	amount paid in September,2005
22	Pay_Amt4	amount paid in September,2005
23	Pay_Amt5	amount paid in September,2005
24	Pay_Amt6	amount paid in September,2005
25	Default_Payment_Next_Month	Amount to be paid next month

Credit [7] is an indenture in which a person designated as borrower receives something of value now and agrees to repay the lender at some date in the future, mostly with interest.

Credit default is not being able to repay or not repaying the debt. This dataset includes 25 attributes of which 24 are features and 1 is a class attribute. With 30000 instances it gives an in depth data view on the subject. The processing performed on this dataset is discussed below.

3.1.2 Data Visualization.

The credit card default dataset is provided in .xls format. This is then converted to .csv format. This is then uploaded onto the data mining tool called WEKA [15], which verifies the data. The dataset has been obtained from a research conducted in Taiwan regarding default customers. Few of the attributes are given in table 2.

Some of the values in the dataset are continuous and others are discrete.

Table 2. Information type in attributes.

Attribute Name	Category	Information
ID	Continue	……
Limit_Bal	Continue	……..
Sex	Discrete	2Value
Education	Discrete	3Value
Marriage	Discrete	2Value
Age	Continue	…….
Pay_0	Discrete	6Value
Pay_2	Discrete	6Value
Pay_3	Discrete	6Value

Pay_4	Discrete	6Value
Pay_5	Discrete	6Value
Pay_6	Discrete	6Value
Bill_Amt1	Continue	…….
Bill_Amt2	Continue	…….
Bill_Amt3	Continue	…….
Bill_Amt4	Continue	…….
Bill_Amt5	Continue	…….
Bill_Amt6	Continue	…….
Pay_Amt1	Continue	…….
Pay_Amt2	Continue	…….
Pay_Amt3	Continue	…….
Pay_Amt4	Continue	…….
Pay_Amt5	Continue	…….
Pay_Amt6	Continue	…….
Default_Payment_ Next_Month	Discrete	2Values

3.1.3 Data Pre-processing

Dataset is obtained in .xls format. This is first converted to .csv format and then passed into the mining tool. The dataset description states that the dataset doesn't miss any data. Feature selection is applied to this dataset.

4. FEATURE SELECTION

Feature selection [8] plays a major role in various spheres of research viz, pattern recognition, statistics, and data mining. Feature selection identifies a subset of the input variables (attributes) by eliminating features with little or no predictive information. Feature selection generally is believed to enhance the accuracy of the resulting classifier and often constructs a model that generalizes better to unseen points. The feature selection algorithms used for this comparison are briefly explained in the following subsections.

4.1 CFC Filtering

It is a supervised feature selection algorithm that adopts the vector space model [9] .Class-Feature-Centroid (CFC) classifier for multi-class, single-label attribute categorization. CFC proposes a novel combination of these indices and employs a denormalized cosine measure to calculate the similarity score between a text vector and a centroid.

4.2 InformationGain Filtering

Information gain is a supervised feature selection algorithm. Information gain (IG) [10] measures the amount of information in bits required for class prediction, if the only information available is the presence of a feature and the corresponding class distribution. The information gain [13] depends on the decrease in entropy after a dataset is split on a selected attribute. Constructing a decision tree aims to identify the attribute that possesses the highest information gain value.

5. CLASSIFICATION

Classification is the process of finding a set of models that describe and distinguish data classes. This is done to achieve the goal of being able to use the model to predict the class whose label is unknown. Some of the algorithms used in our experimental study are briefed in the following sections.

5.1 Naïve Bayes

Naive Bayes classifier [11] is a probabilistic classifier based on the Bayes theorem that assumes that the attributes are independent of each other. **Algorithm** Bayes [13] theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called **class conditional independence.**

5.2 Random Forest

The principle of random forests [12] is to aggregate many binary decision trees built on several bootstrap samples drawn from the learning set. The bootstrap samples are obtained by uniformly drawing 'n' samples among the learning data set with repetition.

5.3 ZeroR

ZeroR is the simplest classification method that relies on the target and gives least importance to the attributes .ZeroR classifier [13] predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. **Algorithm** Construct a frequency table for the target and select its most frequent value. **Predictors Contribution** There is nothing to be said about the contribution of the predictors to the model because ZeroR does not use any of them. **Model Evaluation .**The ZeroR algorithm only predicts the majority class correctly. As mentioned before, ZeroR is only useful for determining a baseline performance for other classification methods.

5.4 SMO

Sequential minimal optimization (**SMO**) [13] is an algorithm that is believed to solve the optimization problem that arises during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. SMO [13] is widely used for training support vector machines and is implemented by the popular libsvm tool. SMO is an iterative algorithm for solving the optimization problem described above.SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multiplier, the smallest possible problem involves two such multipliers.

5.5 Bayesian Network

A Bayesian network B =< N, A, 0 > [14] is a directed acyclic graph (DAG) <N, A> with a conditional probability distribution (CP table) for each node, collectively represented by Θ. Each node n e $\in$ N represents a domain variable, and each arc a $\in$ A between nodes represents a probabilistic dependency [14].In general, a BN can be used to compute the conditional probability of one node, given values assigned to the other nodes; hence, a BN can be used as a classifier that gives the posterior probability distribution of the classification node, given the values of other attributes. When learning Bayesian networks from datasets, we use nodes to represent dataset attributes.

5.6 RANDOM TREE

A random tree is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**.

6. EXPERIMENTAL RESULTS

Evaluation of dataset containing 25 attributes and 30000 is done. The feature selection algorithms have revealed optimal results. Figure 2 shows the feature selection process reducing the number of attributes present in the dataset.

CFS has reduced its attributes from 25 to 6. Similarly, Info Gain has reduced attributes from 25 to 7. The results are discussed below.

6.1 Feature Selection Results

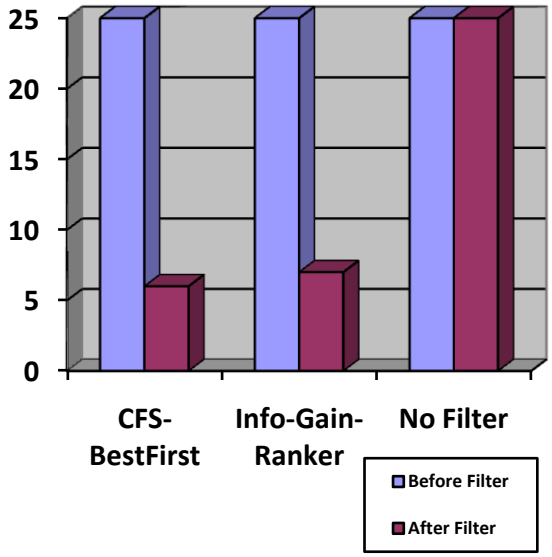

Figure 2.

As evident from Figure 2, CFS and InfoGain have shown good refinement.

Table 3. InfoGain resulting attributes

No	Attribute Name
1	Pay_0
2	Pay_2
3	Pay_3
4	Pay_4
5	Pay_5
6	Pay_6
7	Default Payment next month

Table 4.CFS resulting attributes

No	Attribute Name
1	Pay_0
2	Pay_2
3	Pay_3
4	Pay_4
5	Pay_5
6	Default Payment next month

6.2 Precision And Recall

Precision and recall are two important performance measures. Precision refers to the data that is correctly classified by the classification algorithm. Recall refers to the percentage of data that is relevant to the class [1]. A value of 1.000 depicts a 100% accuracy for both recall and precision.

6.3 Error Rate

Error rate refers to the percentage of the data that is misclassified or wrongly classified. Error rates of different classification algorithms are given in table 5.

Table 5. Error rates comparison for different Classification algorithms

Classification Algorithms	CFS-BestFirst	InfoGain	NoFilter
Bayes Net	0.2315	0.2279	0.2394
Stacking	0.3446	0.3446	0.3446
Naïve Bayes	0.2258	0.2216	0.3703
Random Forest	0.273	0.272	0.2717
Random Tree	0.2732	0.2722	0.2696
SMO	0.1907	0.1908	0.1907
ZeroR	0.3446	0.3446	0.3446
IBK	0.272	0.2717	0.2705

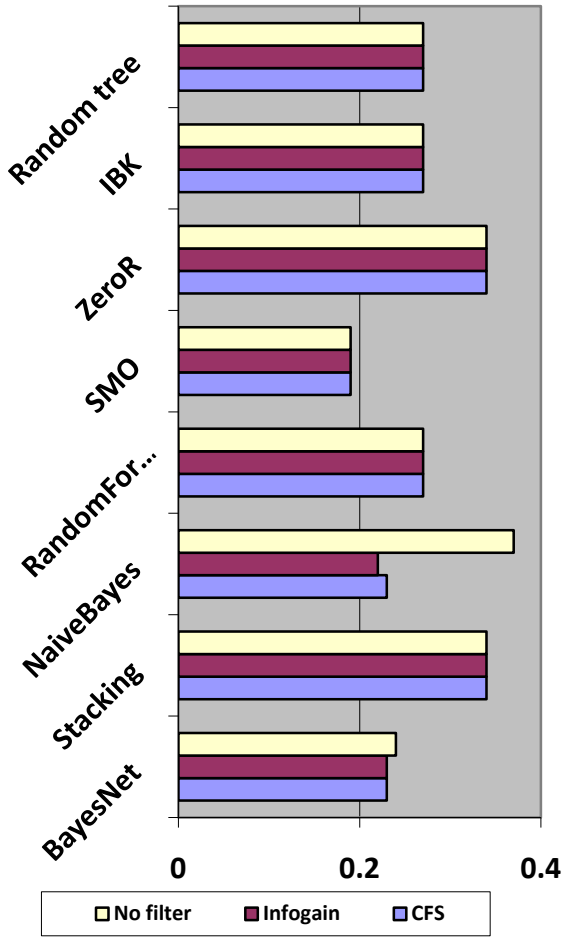

Figure 3: Comparison of error rates in different classifying algorithm with respect to different feature selection.

6.4 Classification Algorithm Results

The classification algorithms have been implemented on the dataset. Following table (Table 6) show the implemented results. WEKA [15] data mining tool has been used to conduct the analysis of the classification algorithms. Efficiency check results of all the classification algorithms used are mentioned in Table 6.

Table 6. Results of all classification algorithms using the data mining tool WEKA [15].

Feature Algorithm	Classification Algorithm	TP	Precision	Recall
CFS-BestFirst	BayesNet	.796	.781	.796
	Stacking	.779	.607	.779
	NaiveBayes	.805	.786	.805
	Random Forest	.816	.798	.816
CFS-	Random Tree	.816	.797	.816
BestFirst	ZeroR	.779	.607	.779
	IBK	.816	.798	.816
	SMO	.809	.793	.809

Feature Algorithm	Classification Algorithm	TP	Precision	Recall
InfoGain	BayesNet	.804	.786	.804
	Stacking	.779	.607	.779
	NaiveBayes	.802	.3784	.802
	Random Forest	.815	.797	.815
	Random Tree	.815	.796	.815
	ZeroR	.779	.779	.779
	IBK	.815	.815	.815
	SMO	.809	.809	.809

BayesNet shows a 80.4 percent accuracy when features selected using InfoGain algorithm are utilized for classification. Similarly, IBK has shown an increasing efficiency on applying feature selection. NaïveBayes has also shown a significant improvement by applying feature selection algorithms (CFS, InfoGain). The classifying algorithm that performed best with the credit card default dataset was Random Tree. Other classifying algorithms like Random Forest and IBK performed equally well.

7. CONCLUSION

Data mining has become pivotal in many fields. They decrease the dependence on man power and also remove manual errors in manipulating the data. The analysis of different classifying algorithms does help future research in this field. First hand implementation of these algorithms would also be aided by analysis of the algorithms. Most algorithms have produced an accuracy of about 80 percent. Application of these algorithms on full scale implementation is cost effective, less error-prone and less time consuming. Application of data mining algorithms in banking sector is gaining momentum. It requires more research and exploration. It has high potential to cater to diversifying needs in the banking sector. RandomTree, RandomForest and IBK which classification algorithms have classified the dataset to good accuracy thus achieving the objective of this study.

8. REFERENCES

[1] Han, Jiawei, and Micheline Kamber. "Data mining: concepts and techniques (the Morgan Kaufmann Series in data management systems)." (2000).

[2] Cios, Krzysztof J., Witold Pedrycz, and Roman W. Swiniarski. *Data mining methods for knowledge*

discovery. Vol. 458. Springer Science & Business Media, 2012.

[3] Dataset from University of California,Irvine available in their online repository http://archive.ics.uci.edu/ml/index.html.

[4] Abbas Kerama and Niloofar Yousefi - A Proposed Classification of Data Mining Techniques in Credit Scoring. Proceedings of the 2011,International conference on industrial engineering and Operations management Kuala Lumpur, Malaysia.

[5] Tudor, Adela Ioana, Adela Bara, and Elena Andrei. "Clustering analysis for credit default probabilities in a retail bank portfolio." *Bucharest Acad Econ Stud, Database Syst J* 3 (2012): 23-30.

[6] Tudor, Adela Ioana, Adela Bâra, and Simona Vasilica Oprea. "Comparative analysis of data mining methods for predicting credit default probabilities in a retail bank portfolio." *WSEAS International Conference. Proceedings. Recent Advances in Computer Engineering Series*. No. 7. WSEAS, 2012.

[7] Investopedia is an online encyclopedia relating to investments. http://www.investopedia.com/terms/c/credit.asp

[8] Kim, Yong, W. Nick Street, and Filippo Menczer. "Feature selection in data mining." *Data mining: opportunities and challenges* 3.9 (2003): 80-105.

[9] Guan, Hu, Jingyu Zhou, and Minyi Guo. "A class-feature-centroid classifier for text categorization." *Proceedings of the 18th international conference on World wide web*. ACM, 2009.

[10] Roobaert, Danny, Grigoris Karakoulas, and Nitesh V. Chawla. "Information gain, correlation and support vector machines." *Feature Extraction*. Springer Berlin Heidelberg, 2006. 463-470.

[11] Jacob, Shomona Gracia, and R. Geetha Ramani. "Discovery of knowledge patterns in clinical data through data mining algorithms: multi-class categorization of breast tissue data." *International Journal of Computer Applications (IJCA)* 32.7 (2011): 46-53.

[12] Genuer, Robin, et al. "Random Forests based feature selection for decoding fMRI data." *Proceedings Compstat*. No. 267. 2010.

[13] Nasa, Chitra. "Evaluation of different classification techniques for web data."*International Journal of Computer Applications* 52.9 (2012).

[14] Cheng, Jie, and Russell Greiner. "Comparing Bayesian network classifiers."*Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999.

[15] WEKA- Weka is a collection of machine learning algorithms for data mining tasks.http://www.cs.waikato.ac.nz/ml/weka/

[16] Watanabe, Carolina YV, et al. "SACMiner: A new classification method based on statistical association rules to mine medical images." *Enterprise Information Systems*. Springer Berlin Heidelberg, 2010. 249-263.