A Comparative Study and enhancement of classification techniques using Principal Component Analysis for credit card dataset

Abhishek Agarwal
B.Tech
Computer Science & Engineering.
Inderprastha Engineering College
Ghaziabad, Uttar Pradesh.

Amit Rana
B.Tech
Computer Science & Engineering.
Inderprastha Engineering College
Ghaziabad, Uttar Pradesh.

Karan Gupta
B.Tech
Computer Science & Engineering.
Inderprastha Engineering College
Ghaziabad, Uttar Pradesh.

Neeta Verma
Associate Professor
Computer Sc. & Engineering.
Inderprastha Engineering College
Ghaziabad, Uttar Pradesh.

***Abstract*: The following research reveals the significance of modified classification in estimating new trends. Rigorous evaluation of different classification algorithms viz. Logistic Regression, Decision Tree, K-Nearest Neighbor and Naive Bayesian is explored in this paper. These findings forecast the finest techniques for discovery of potential defaulters which can be adapted by banking institutions. Our main motive is to compare the performance measures between original dataset and original dataset on which principal component is applied. The reason to use the principal component was to evaluate its impact on the performance of the algorithms used while dealing with the dataset. Different algorithms can be compared on the basis of various criterions such as Accuracy, Precision, F1-Score, Recall, ROC. Successful contrast between these attributes would yields a efficient model for the given dataset .Logistic regression is then found to be the most efficient method for this particular dataset.**

General Terms
Decision tree, K-Nearest Neighbor, Logistic Regression, Naive Bayesian, Principle Component Analysis.

I. INTRODUCTION

As a result of current rising trend of using credit card more frequently as a payment mode. There is a sharp elevation in the number of defaults in the process of paying back said amount. This not only affects the status of the credit card holder of the current bank but also in general with the other banks. In-fact this can lead to much stricter criterion for loan approval, new credit card approval and other credit based services in general.

During the process of signing of credit card, a customer must adhere to the set of rules laid by the bank. A prominent condition of this includes the time limit for the repayment of the credit taken. It is advised that one should not miss monthly credit card bill thereafter. However, even upon missing such deadlines you would be subjected to additional allowable time within which you must fulfill the payments. If there is a failure of accommodation of payment in this grace period then it would lead to a default. The time period is usually of 6 months. The annual interest rate varies from 30% to 45%. Even after a single day following the penalty date, an individual is subjected to the full outstanding amount. This result attracts all kinds of unfavorable behavior such as blacklisting by the credit Bureaus, legal action by law firms associated with the bank, higher rate interest charge. Resolving a credit card default hence is tedious and ever slandering methods of payment.

II. RELATED WORK

A lot of extensive research on credit card default dataset has been performed over the years. Ajay et al. have done a research work and explained that both correlation feature subset selection and information gain feature selection methods yield the most handy features for prediction. Classification algorithms application implemented on full scale is cost effective, less error prone and less time consuming. Random Forest method is best in forecasting credit card defaulters. Random forest and IBK algorithms gave good accuracy as compared to other algorithms such as naive Bayes, random forest, zero R, Bayesian network, Random tree. IBK was most precise with the precision value being 0.815 [1]. S. Saini et al. used various performance measures such as accuracy, execution time, correctly classified instances, incorrectly classified instances and error rate. They used multi-layer perceptron with 10 folds cross validation method with accuracy close to 99.75%. When they used j48 method, it was 100% accurate. [2] Yeh et al. compared the predictive accuracy of probabilities of 6 data mining methods- K-NN, logistics regression, Naive Bayes, classification trees, ANN, discriminant analysis. Among the 6 data mining techniques, ANN was the single technique that could finely estimate the original probability of the default. They compared the performance of the classifications that predict the accuracy among the 6 technique. In the predictive accuracy of probability of defaults, ANN showed the most optimal performance on probability of defaults, ANN showed the most optimal performance on regression intercept (0.0175) &regression coefficient (0.9971), the predictive default proba-

978-1-7281-4097-1/20/$31.00 ©2020 IEEE

bility which is the output by ANN is the single technique that can be employed to represent original probability of default. In the training set discriminant analysis performance bad, KNN perform the best with error rate of 0.18. During testing, discriminant analysis performs worst this time also and K-NN perform the best with error rate of 0.16. [3] A.Verma observed that on the dataset the average precision, weighted average recall, the weighted average recall and the weighted average F1 score have the highest level of performance on followed by multilayer Neural Network. Comparable performance was seen by decision tree with the worst performance seen in Naive Bayes & the simple K-NN. [4] Reddy et al. compares various classification algorithms, the results are as following, the ID3 algorithms gave the highest accuracy but had a long searching time and took more memory than c4.5 algorithms to large program execution. On the other hand, naive Bayes provide the balanced approach giving a good performance O(n) time performance, while it requires a very large number of records to obtain good results [5]

III. DATASET

The dataset that was adapted has documented the transactions of over 30,000 customers over a span of 6 months from April 2005 to September 2005. The dataset is a multivariate with both integer and real values and doesn't contain any missing value. It has 24 attributes and one more additional attribute as response variable[6].

The dataset is obtained from UCI Repository credit card defaulter. In the first table the attributes are as follows Pay_0 then Pay_2 to Pay_6 and it shows repayment status from April to September 2005.

Table 1: Attribute Pay and their description

S.No	Attribute Name	Description
1	Pay_0	The status of repayment in September, 2005
2	Pay_2	the status of repayment in August, 2005
3	Pay_3	the status of repayment in July, 2005
4	Pay_4	the status of repayment in June, 2005
5	Pay_5	the status of repayment in May, 2005
6	Pay_6	the status of repayment in April, 2005

The second table shows the status of the bill amount from April to September in the year 2005 and attributes names are fromBill_Amt1 to Bill_Amt6.

Table 2: Attribute Bill_Amt and their description

S.No	Attribute Name	Description
1	Bill_Amt1	amount of bill statement in September,2005
2	Bill_Amt2	amount of bill statement in August,2005
3	Bill_Amt3	amount of bill statement in July,2005
4	Bill_Amt4	amount of bill statement in June,2005
5	Bill_Amt5	amount of bill statement in May,2005
6	Bill_Amt6	amount of bill statement in April,2005

The third table shows the status of the amount paid from April to September in the year 2005 and attributes names are fromPay_Amt1to Pay_Amt6.

Table 3: Attribute Pay_Amt and their description

S.No	Attribute Name	Description
1	Pay_Amt1	amount paid in September,2005
2	Pay_Amt2	amount paid in August,2005
3	Pay_Amt3	amount paid in July,2005
4	Pay_Amt4	amount paid in June,2005
5	Pay_Amt5	amount paid in May,2005
6	Pay_Amt6	amount paid in April,2005

The fourth table shows the different attributes along with their description. The attributes are ID, Limit_balance and Education etc. of the dataset

Table 4: Other attributes and their description

S.No	Attribute Name	Description
1	ID	ID of the user
2	Limit_Balance	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	Default_Payment_Next_ Month	Amount to be paid next month

This table shows the types or category of the attributes used in the dataset. Some of the values in the dataset are continuous and others are discrete.

Table 5: Attributes and their types

Attribute Name	Type of Attribute
ID	Continuous
Limit Balance	Continuous
Sex	Discrete
Education	Discrete
Marriage	Discrete
Age	Continuous
Pay_0	Discrete
Pay_2	Discrete
Pay_3	Discrete
Pay_4	Discrete
Pay_5	Discrete
Pay_6	Discrete
Bill_Amt1	Continuous
Bill_Amt2	Continuous

Bill_Amt3	Continuous
Bill_Amt4	Continuous
Bill_Amt5	Continuous
Bill_Amt6	Continuous
Pay_Amt1	Continuous
Pay_Amt2	Continuous
Pay_Amt3	Continuous
Pay_Amt4	Continuous
Pay_Amt5	Continuous
Pay_Amt6	Continuous
Default_Payment_ Next_Month	Discrete

IV. METHODOLOGY

PCA - Principal component analysis [7] is a statistical tool that comes in handy when dealing with a lot of potentially correlated attributes. It reinstates said variables as a set of linearly uncorrelated variables using orthogonal transformation such that the prime component displays the largest possible variance with each subsequent term having the highest variance possible while still adhering to the orthogonality with its preceding component.

ACCURACY – The compactness of the measured value to familiar value. In this case the number of true positives, true negative, false positive and false negative help in defining accuracy as following [8]

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

PRECISION – For repeated measurements under continuous conditions, the level to which data shows the identical result is known as precision. It is due to closeness of the measured values to each other. [9]

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

ROC – At the different classification threshold, it is between true positive rate and false positive rate. It is used to conceptualize the performance of binary classifier [10].

4.5 F1 Score - It is defined as the harmonic mean of the precision and recall and helps in getting a clear indication of the precision and accuracy of a model.[11]

$$\text{F1 score} = 2 \text{ x } \frac{PrecisionxRecall}{Precision+Recall} \qquad (3)$$

RECALL - This is also expressed as the division of the true positives upon the sum of the true positives and the false negatives.[12]

$$\text{Recall} = \frac{TP}{TP+FP} \qquad (4)$$

Logistic Regression - A Logistic Regression [13] is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression comes handy in figuring out the relationship between a dependent binary variable and independent variables.

$$\text{Log}\left(\frac{Pi}{1-Pi}\right) = \beta_0 + \beta_1X_1 + \beta_2X_{2....}\beta_kX_k \qquad (5)$$

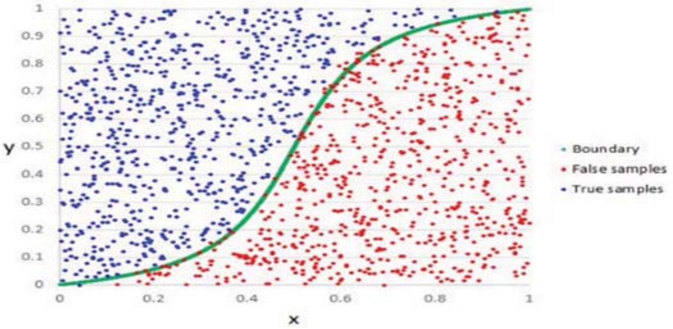

Fig -1 Logistic Regression

Decision Tree - A decision tree [14] is a tree whose interior nodes can be considered as tests (on operant data patterns) and whose terminal nodes can take as categories (of such patter). Decision trees contain tree like structure and are constructed by using an algorithmic approach to identify the ways to split a data set depending upon given conditions. It maps all possible outcomes. It starts with a single node and branches into possible outcomes. Then each outcome connects to additional nodes which branches off into another outcome. In this way it becomes tree like structure. There are many decision tree algorithms and it is necessary to find which algorithm is best for our dataset.

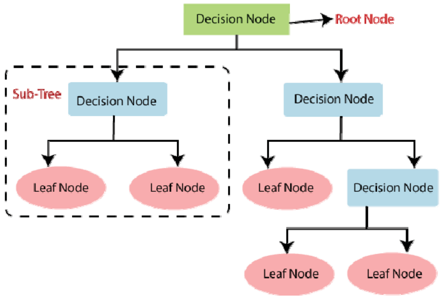

Fig-2 Decision Tree

Naive Bayes: -Naive bayes[15] is a robust classifier in this sense and has the least storage requisites and comparatively speedy training, thus being applicable to time-storage dependant fields which ranges from voluntarily assembling internet pages into certain types and apply spam filtering. Naive bayes theorem provides a way of finding the posterior probability, P (c|x), from P(c), P(x), and P(x|c). This classifier automatically defaults to the observation that the affect of the predictor(x) on a class(c) provided is clearly unaffected by the values of other predictors. The above assumption is termed as class conditional independence. P(c|x) is termed as posterior probability of the class (target) given predictor (attribute). P(c) is the prior probability of class. P(x|c) represents the chances of the predictor given class. P(x) is the probability of a predictor that comes in prior

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)} \qquad (6)$$

K-Nearest Neighbor– K-NN[16] a type of learning based on instances, where the function is only approximated locally and all computation is deferred until classification. This algorithm can be quite handy to distribute the weights to the partnership such that the nearer of the neighbors have more partnership to the average than the ones that are further away. For example, a common weighting scheme helps in providing every neighbor a weight of 1/d, where d is the distance towards the neighbor. The neighbors are selected from a set of objects for which the class is identifiable.

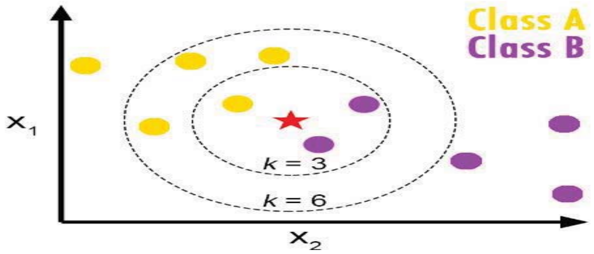

Fig -3 K – Nearest Neighbor

V. EXPERIMENTAL RESULTS

Before applying Principal Component Analysis, it was found in the test set that the logistics regression outperformed Decision tree, K-Nearest Neighbor, Naïve Bayesian by a margin of at least 0.019 in terms of accuracy while was it subpar in terms of precision and F1 score with comparable values of ROC and Recall. Decision tree on the other hand showed good numbers in terms of ROC while remaining comparable in every other aspect. We can also observe that the values of precision, accuracy and recall before applying PCA weren't as good as after applying it. For instance, the value for recall in case of logistic regression has seen a steady increase of 0.225 while in case of K-Nearest Neighbor, there has been an increase of almost 0.224 after applying Principal Component Analysis. Finally, we see the ROC value for all of the algorithms increase by at least 0.050. In case of training set before applying PCA, it was found that decision tree performed best with all performance values equal to 1. In training K-Nearest Neighbor performed best in terms of accuracy with value equal to 0.800 and in terms of precision also with value equal to 0.614.

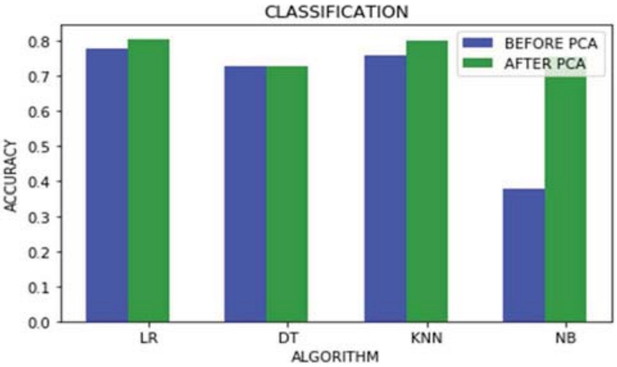

Fig. 4: Comparison of accuracy before and after PCA

Table 6: Values before applying principal component analysis

Techniques	Accuracy	Precision	Recall	F1 Score	ROC
Logistic Regression	0.777	0	0	0	0.5
Decision Tree	0.728	0.391	0.401	0.396	0.611
k-Nearest Neighbour	0.758	0.89	0.155	0.221	0.542
Naïve Bayesian	0.377	0.249	0.895	0.389	0.562

Techniques	**ACCURACY**	**PRECISION**	**RECALL**	**F1 SCORE**	**ROC**
Logistic Regression	0.806	0.699	0.225	0.341	0.598
Decision Tree	0.729	0.392	0.396	0.394	0.610
KNN	0.801	0.592	0.339	0.431	0.636
Naïve Bayesian	0.754	0.458	0.596	0.518	0.697

Table 7: Values after applying principal component analysis

After applying Principal Component Analysis in the test set, we observe that logistic regression had the best accuracy and precision as compared to other elements. Naive Bayes had the best recall, f1-score and roc value of 0.596, 0.518 and 0.607 respectively. This clearly illustrates the benefit of using Principal Component Analysis in our use case. In the training set, after applying Principle component analysis ,it was observed that decision tree performed best in terms of performance parameters with value equal to 1.Logistic Regression performed best in terms of precision while K-Nearest Neighbor performed best in terms of accuracy in the training set.

Figure 4 above clearly shows that the values for accuracy of every algorithm has seen a minimum rise except in case of Naïve Bayes which has seen a significant rise (an increase of 0.377) after applying PCA.

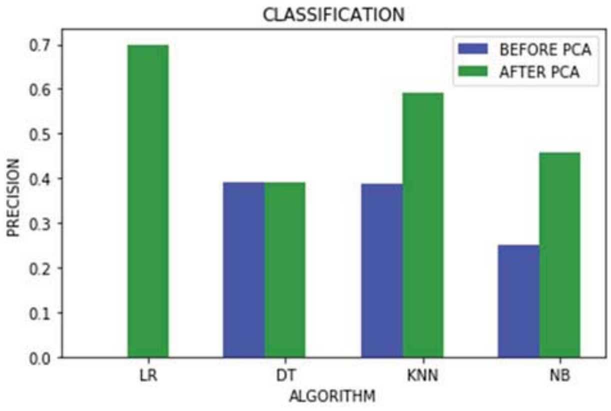

Fig. 5 Comparison of precision before and after PCA

Figure 5 shows that the precision has seen significant improvement in case of Logistic regression with comparable increase in KNN and Naïve Bayes. On the other hand, the increase in decision tree was minimum (an increase of 0.001).

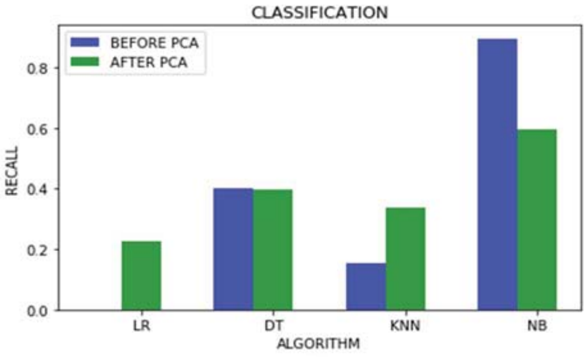

Fig. 6: Comparison of recall before and after PCA

Figure 6 illustrates the decrease in performance in case of Naïve Bayes while there was significant increase in logistic regression and appreciable increase in case of KNN, while Decision tree showed minimal change in its case.

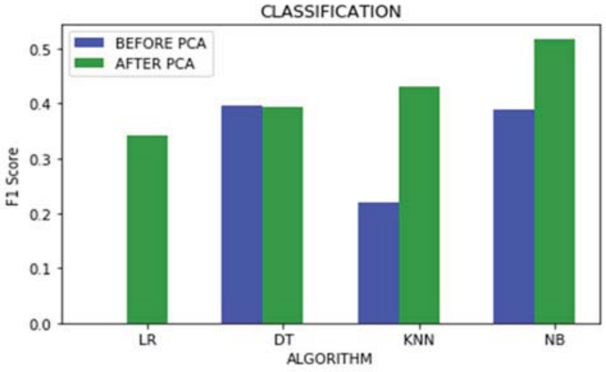

Fig. 7: Comparison of f1 score before and after PCA

Figure 7 above shows significant increase in case of Logistic Regression with comparable increase in case of Naïve Bayes and K Nearest Neighbor and finally decrease in performance in case of Decision tree.

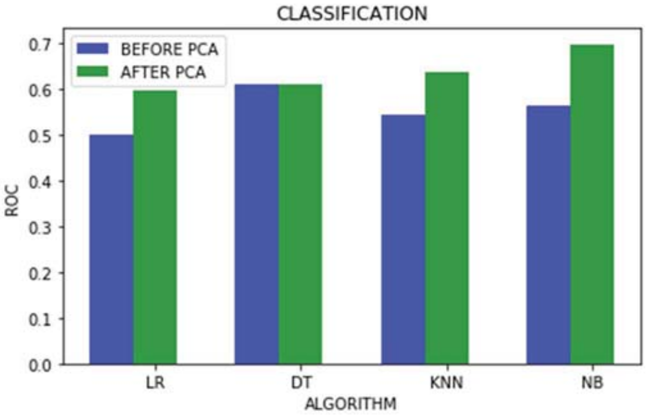

Fig. 8: Comparison of ROC before and after PCA

Figure 8 shows little difference in case of decision tree after applying PCA and also comparable increases in case of Logistic regression, K-NN and Naïve Bayes. There is a increase in ROC value after applying the principal component analysis

VI. CONCLUSION :-

These results have become pivotal for banking institutions. In our work various performance measures were compared and it was found that accuracy for logistic regression was best in both the cases. The results showed that decision tree performance was not much affected by the principal component analysis. Precision value was not much affected for logistic regression as compared to other classification methods. The other performance measures such as ROC, F1-Score showed good results for naïve Bayesian. K-Nearest neighbor showed acceptable performance in terms of recall.

REFERENCES :-

[1] Ajay, A.Venkatesh, S.G.Jacob,"Prediction of credit card defaulters:acomparative study on performance of classifiers", International Journal of Computer Applications (0975 – 8887) Volume 145 – No.7, July 2016

[2] S.Saini, A.Dhanbbar and Dr. K.Solanki, "Comparative study of classification algorithm using weka", IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 08, Issue 10 (October. 2018), ||V (II) || PP 29-40

[3] I-Cheng Yeh, Che-Hui Lien, "Compared the predictive accuracy of probablity of 6 datamining methods- K-NN, logistics regression, naive bayes, classification trees, ANN, discriminant analysis" Elsevier Journal, Volume 6 issue, Feb 2007, Page 1,7.

[4] A.Verma, "Study and Evaluation of classification algorithms in data mining*" International Research Journal of Engineering and Technology*, Volume 5 Issue 8, Aug 2018 ,page-1,11.

[5] N.C.Reddy, K.S.Basand and Mounika, "Classification algorithms on data mining: a study compares various classification algorithms" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume13, Number 8 (2017), pp. 2135-2142

[6] Default of credit card clients dataset "
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[7] Principal Component Analysis (PCA)"https://www.techopedia.com/definition/32509/principal-component-analysis-pca

[8]"Classification:Accuracy"https://developers.google.com/machinelearning/crashcourse/classification/accuracy

[9]"Accuracy,Precision,Recall,F1?"https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[10]"Classification:ROCCurveandAUC"https://developers.google.com/machinelearning/crashcourse/classification/roc-and-auc

[11]" What is the F Score?" https://deepai.org/machine-learning-glossary-and-terms/f-score

[12]"Classification:PrecisionandRecall"https://developers.google.com/machinelearning/crashcourse/classification/precision-and-recall

[13] Chai-Ying Joanne Peng, KukLida Lee, Gary M. Ingersoli, "An Introduction to Logistic Regression Analysis and Reporting", the Journal of Educational Research, Vol. 96(No. 1), September/October 2002

[14] H. Sharma, S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining",International Journal of Science and Research (IJSR) 5(4) · April 2016

[15] P. Kaviani, S. Dhotre , "Short Survey on Naive Bayes Algorithm", International Journal of Advance Research in Computer Science and Management 04(11) · November 2017

[16] Yun-lei Cai, Duo Ji ,Dong-fengCa, "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor", NTCIR-8 Workshop Meeting, June 15–18, 2010

[17] S. Rajora*et al.*, "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bangalore, India, 2018, pp. 1958-1963.

[18] B. Emil Richard Singh and E. Sivasankar, "Risk Analysis in Electronic Payments and Settlement System Using Dimensionality Reduction Techniques," 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, 2018, pp. 14-19.

[19] C. S. Michael Strumpf, "Detecting loan defaults at an early stage using models of machine intelligence," Analytics India Magazine, pp. May–11, 2015.

[20] H. Abdi and L. J. Williams, "Principal component analysis," Wiley interdisciplinary reviews: computational statistics, vol. 2, no. 4, pp. 433–459, 2010.

[21]J.GalindoandP.Tamayo,"Creditriskassessmentusingstatistical and machine learning: basic methodology and risk modeling applications," Computational Economics, vol. 15, no. 1, pp. 107–143, 2000.

[22] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.