Predicting the Credit Defaulters using Machine Learning Techniques

[1]Nagaseshu Kurapati , [2]Priyanka Kumari Bhansali
[1]M.Tech Student, [,2]Research scholar
[1,2]Department of Computer Science and Systems Engineering
[1,2]Andhra University College of Engineering, Visakhapatnam, AP, India
k.nagaseshu026@gmail.com

Abstract

In today's world, people are heavily depending upon financial institutions to take loans for different purposes. Making Applications from people are increasing day-to-day. Financial institutions cannot approve all these applications due to their reliability. Financial institutions may face huge capital loss if they sanctioned the loan without having any prior assessment of default risk. In addition, the number of transactions in the banking sector is rapidly increasing and huge data volumes are available which represent the customer's behavior and the risks around loan are increased. Financial institutions always need a more accurate predictive modeling system for many issues. Predicting credit defaulters is a crucial task for the banking industry. Machine learning is a promising area to extract patterns from the huge volumes of the data. This paper used different machine learning algorithms to build an efficient model to improve the accuracy of the defaulter's prediction.

Keywords: *Machine Learning, Data mining, Random Forest, Decision Tree, Sampling, PCA.*

1. Introduction

Acquiring loans for different purposes such as the Home loan, Education Loan, car loan, Business Loans etc., has become part of our day-to-day life from financial institutions like banks and credit unions. However, some people are unable to properly gauge the amount of loan that they can afford. In some cases, people are unable to pay back bulk loans due to the sudden financial crisis while some try to scam money out of the banks. Prior assessment of defaulters, which involved in a loan application, is one of the most important concerns of the banks for survival in the highly competitive market and for profitability. Everyday Financial institutions receive a number of loan applications from their customers and other people. Not everyone gets approved by the banks or financial institutions. Most of the banks use their own credit scoring models and risk assessment techniques in order to analyze the loan application and to make decisions whether to approve the application or not. In spite of this, financial institutions facing huge capital loses with the default of loans by the people. In this project, Machine learning algorithms will be used to study the historical credit data to extract patterns from it, which would help in predicting the likely defaulters, thereby helping the financial institutions for making better decisions in the future.

2.Literature Review

Researchers have proposed many models to predict credit defaulters some of them are:

Sudhamathy G and Jothi Venkateswaran designed the decision tree based classification model to predict the defaulters. This paper tested and trained the model using the data set available in the UCI repository. Prior to building the model, the dataset is pre-processed, reduced and made ready to provide efficient predictions[1]. The objective of the Arutjothi and Senthamarai is to create a credit scoring model for credit data. Various machine learning techniques are used to develop the financial credit scoring model. This paper used a K-Nearest Neighbor (K-NN) classifier to build the model for credit data[2][9]. Six data mining techniques (FLDA, Naïve Bayes, J48, Logistic Regression, MLP, and IBK) are applied to the dataset to compare the performance of the algorithms in predicting the credit defaulters. The motive of this research is to compare the predictive accuracy of the customer's default payments using different data mining techniques[3]. Sudhamathy aims to design a model and prototype using a data set available in the UCI repository. The model is a decision tree based classification model that uses the functions available in the R Package. Random Forest, Logistic Regression, SVM classification algorithms are studied and analyzed the bank credit data set, and compared these models on five model effect evaluation statistics of Accuracy, Recall, precision, F1-score, and ROC area[5]. Ehsan Kamalloo and Mohammad Saniee Abadeh designed two forms of memory: simple memory and k-layer memory. This model used immune memory to remember good B cells during the cloning process. Two real-world credit data sets in the UCI machine learning repository are selected as experimental data to show the accuracy of the proposed classifier.[6]. This paper proposes two credit scoring models using data mining techniques to support loan decisions for the Jordanian commercial banks. Loan application evaluation would improve credit decision effectiveness and control loan office tasks, as well as save analysis time and cost[8][10].

SHIYI CHEN and W. K. HÄRDLE proposed the Support Vector Machine (SVM) to predict the default risk of German firms. In all tests performed in this paper, the nonlinear model classified by SVM exceeds the benchmark logit model, based on the same predictors, in terms of the performance metric, AR[7]. José Francisco Martínez Sanche and Gilberto Pérez Lechuga presented the evaluation of a credit scoring system in terms of cost-efficiency for savings and loan institutions in specific for SOFIPO's and in terms of cost-benefit to the service provider assessment of loan applications[11]. Hussein A. Abdou and Marc D applied the logistic regression (LR), Classification and Regression Tree (CART) and Cascade Correlation Neural Network (CCNN) in building knowledge-based scoring models. The model used ROC curves and Gini coefficients as evaluation criteria and the Kolmogorov-Smirnov curve as a robustness test To compare various models performances[12].

3. Proposed Methodology

Data Collection

The proposed system implementing algorithms on the historical credit data which is collected from the UCI repository. The dataset contains 21 columns of 1000 records, which describes different features of the dataset.

Data preprocessing

Data preprocessing is an important task to be done prior to analysis to get the data ready for analysis. As good data can only provide better results, data preprocessing

becomes necessary prior to analysis. In data preprocessing, the proposed system performs data cleaning, data imputation, data normalization, and transformation. Data cleaning process removes null values and redundant attributes from the dataset.

Implementation

The Proposed model applies the sampling technique(PCA) on the preprocessed dataset to balance it. On this sampled data, The proposed system implements the Machine Learning Algorithms to check which algorithm suits better i.e., which algorithm is suitable for prediction. This system also compares the accuracy of algorithms before and after feature selection to select the best algorithm that predicts the defaulters effectively. The architecture of the proposed system is given in fig.1.

.

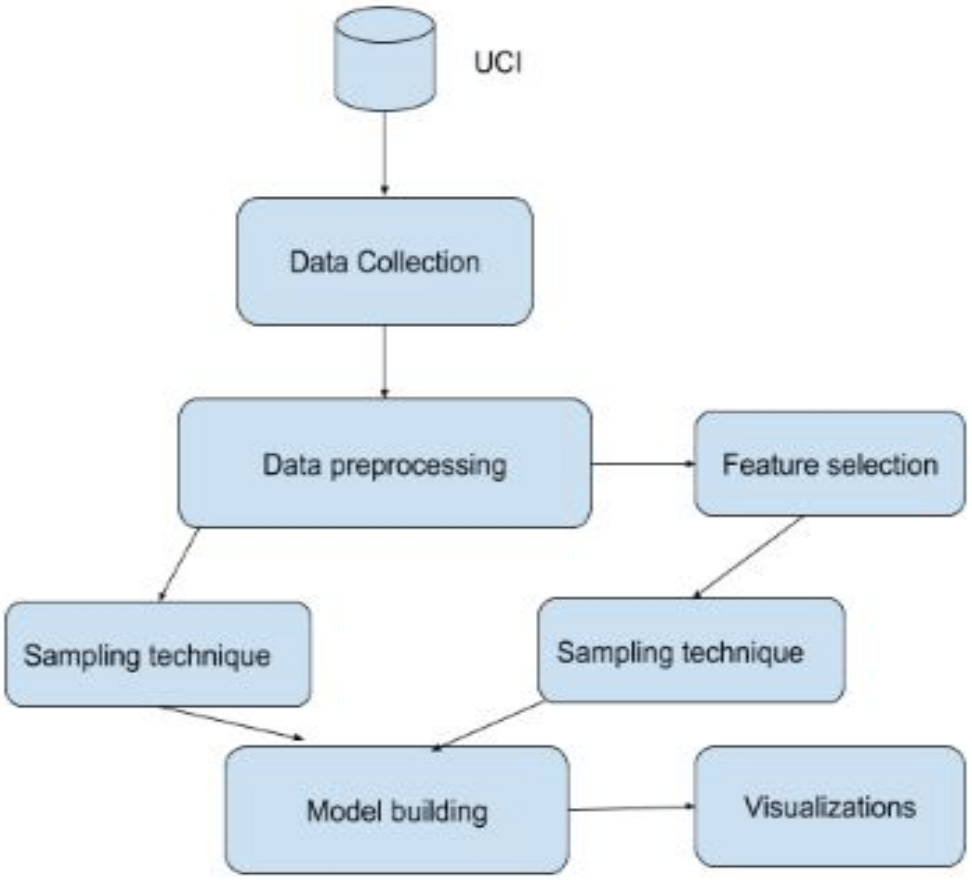

Fig.1 Proposed model

4. Results

Machine Learning algorithms are implemented on the Credit historical data which is collected from the UCI repository. The proposed model applied the algorithms on the data before and after feature selection to compare the performance of the algorithms. It is evident from the result that feature selection has improved the performance algorithms in predicting the defaulters. The performance of algorithms before feature selection is given in table[1].

Table.1 Accuracy of algorithms before feature selection

Model	Accuracy	Precision	Recall
Extra tree classifier	0.94	0.99	0.90
Decision Tree	0.86	0.98	0.73
Random Forest	0.95	1.00	0.89
Gradient Boosting	0.9	0.99	0.81

To select the important features from the whole set of data proposed system used the Principle component analysis. The proposed selected features are shown in Fig.1.

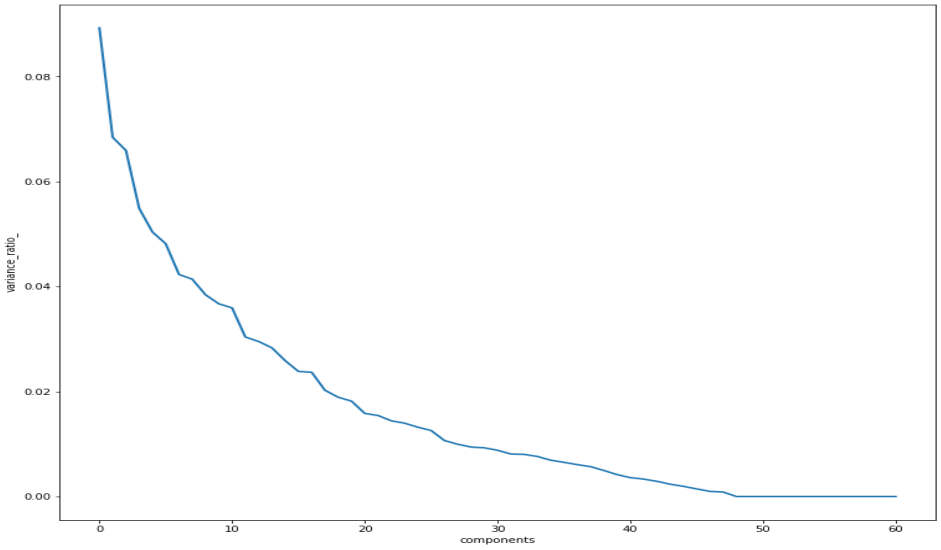

Fig.2 Feature selection using PCA

Performance of the algorithms after feature selection is given in table[2].

Table.2 Performance of Algorithms after feature selection

Model	Accuracy	Precision	Recall
Extra tree classifier	0.95	0.98	0.91
Decision Tree	0.89	1.00	0.77
Random Forest	0.97	1.00	0.94
Gradient Boosting	0.91	1.00	0.80

```
cnf_matrix: [[63  4]
 [ 0 73]]
              precision    recall  f1-score   support

           1       1.00      0.94      0.97        67
           2       0.95      1.00      0.97        73

   micro avg       0.97      0.97      0.97       140
   macro avg       0.97      0.97      0.97       140
weighted avg       0.97      0.97      0.97       140

TestAccuracy : 0.9714285714285714
Precision: 1.0
Recall: 0.9402985074626866
F1 score: 0.9713699633699633
```

Fig.3 Classification report of Random forest

5. Conclusion

In this paper, in order to predict credit defaulters, historical credit data collected from the UCI repository. The collected data is not in balanced mode. So to balance the dataset, a sampling technique is applied to it. In this work, we used different Machine Learning algorithms for defaulters predictions. It is evident from the results that, Random Forest algorithm outperforming the other models in identifying the defaulters. Credit defaulters prediction can be useful to many people if its prediction is more accurate, which can be attained by applying Deep Neural Network based algorithms like, RNN etc.

6. References

[1] Sudhamathy G, Jothi Venkateswaran C, "Analytics Using R for Predicting Credit Defaulters," in International Conference on Advances in Computer Applications, 2016

[2] G Arutjothi, C Senthamarai, "Prediction of loan status in the commercial bank using machine learning classifier," in International Conference on Intelligent Sustainable Systems, 2017

[3] Dr. Maruf Pasha, Meherwar Fatima, "Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters", 2017

[4] Sudhamathy G, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R" 2016

[5] Li Ying, " Research on bank credit default prediction based on data mining algorithm," 2018.

[6] Ehsan Kamalloo and Mohammad Saniee Abadeh, "Credit Risk Prediction Using Fuzzy Immune Learning, " 2014.

[7] SHIYI CHEN, W. K. HÄRDLE, "Modeling default risk with Support Vector Machines, " 2011.

[8] Yujie Liu, Song Ma, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach, " in IITA International Conference on Services Science, Management and Engineering, 2009

[9] Mahesh Mardolkar, "Loan Defaulter's Application R Programming," 2016.

[10] Hussain Ali Bekhet, Shorouq Fathi Kamel Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural the scoring approach, " 2014.

[11] José Francisco Martínez Sánchez, Gilberto Pérez Lechuga, "Assessment of a credit scoring system for popular bank savings and credit, " 2016.

[12] Hussein A. Abdou, Marc D. Dongmo Tsafack, "predicting creditworthiness in retail banking with limited scoring data," 2016.

[13] G. Arutjothi, G. Arutjothi, "A Prediction Model for Credit Defaulters using Ensemble Learning Classifiers, " 2017.