International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Machine Learning Based Suspicion of Customer Detention in Banking with Diverse Solver Neighbors and Kernels

M. Shyamala Devi, Jyotikinkar Saharia, Shubham Kumar, Aayushi Chansoriya, Prashant Yadav

Abstract*: *In the current moving technological business sector, the amount spent for attaching the new customer is highly expensive and time consuming process than adopting some methods to hold and retain the existing customers. So the business sector is in need to make a research on with holding the existing customers by using the current technology. The methods to make the retention of the existing customers with high reliablility are a challenging task. With this view, we focus on predicting the customer churn for the banking application. This paper uses the customer churn bank modeling data set extracted from UCI Machine Learning Repository. The anaconda Navigator IDE along with Spyder is used for implementing the Python code. Our contribution is folded is folded in three ways. First, the data preprocessing is done and the relationship between the attributes are identified. Second, the data set is reduced with the principal component analysis to form the 2 component feature reduced dataset. Third, the raw dataset and 2 component PCA reduced dataset is fitted to various solvers of logistic regression classifiers and the performance is analyzed with the confusion matrix. Fourth, the raw dataset and 2 component PCA reduced dataset is fitted to various neighboring algorithms of K-Nearest Neighbors classifiers and the performance is analyzed with the confusion matrix. Fifth, the raw dataset and 2 component PCA reduced dataset is fitted to various kernels of Support Vector Machine classifiers and the performance is analyzed with the confusion matrix. The implementation is carried out with python code using Anaconda Navigator. Experimental results shows that, the rbf kernel of Support vector machine classifier is effective with the accuracy of 85.8% before applying PCA and accuracy of 80.9% after applying PCA compared to other classifiers.

Index Terms*: *Machine Learning, Churn, Classification, accuracy, precision, recall and f-score.

Manuscript published on November 30, 2019.
* Correspondence Author
M. Shyamala Devi*, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
Jyotikinkar Saharia, Third Year B. Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
Shubham Kumar, Third Year B. Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
Aayushi Chansoriya, Third Year B. Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
Prashant Yadav, Second Year B. Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

I. INTRODUCTION

The customer feedback analysis is the major concern for the business industries to keep track of the existing customers' ratio for the retention analysis. The existing customers have the power of passing their feedback and subscription to their friends which enlarges the business networks. The retention of the existing is more essential for increasing the new service request from the new customers.

The paper is organized in such a way that Section 2 deals with the related works. Section 3 discuss about the proposed work followed by the implementation and Performance Analysis in Section 4. The paper is concluded with Section 5.

II. RELATED WORK

A. Literature Review

The customer churn is one of the major challenges in the telecommunication industry sector. The attraction and the retention process of the new customers towards the business market is highly expensive than the retention of the existing customers. A broad survey on analysis of churn prediction is done by evaluating the data by using the classifiers for the telecommunication company in Macedonia is done [1].

Decision support systems for the churn prediction can be done using the data mining and knowledge discovery techniques. The revenue loss is mainly due to the loss of existing customers. A decision support system is design to manage the customer retention with the scalable company growth [2]. The churn management for the mobile communication is very important to with stand and hold the existing customers before moving to the other services. The churn prediction is done by data mining techniques by empirical evaluation using decision tree method. The basic information of the customer service log is also used for the customer churn prediction [3]. The major challenge in any business sector is to reduce the service request from their existing and new customers. This can decrease the operational cost of the business. The customer churn analysis for the Brazilian telecom sector is done using classification techniques by designing the dataset as a binary classification task [4].

The analysis of the same resembling features of the existing customers is the challenging task for the business decision makers for analyzing the customer churn. The customer churn prediction is done by dividing the dataset into low integrity dataset and high integrity dataset [5]. The feature selection and extraction methods are reviewed from these papers [6]-[24]

III. PROPOSED WORK

In our proposed work, the customer churn bank modelling data set is subjected to predict the customer churn analysis. Our implementation in this paper is folded in six ways.

(i) First, the data preprocessing is done and the relationship between the attributes are identified.

(ii) Second, the data set is reduced with the principal component analysis to form the 2 component feature reduced dataset.

(iii) Third, the raw dataset and 2 component PCA reduced dataset is fitted to various solvers of logistic regression classifiers like liblinear, lbfgs, newtoncg, sag, saga and warn and their performance is analyzed with the confusion matrix.

(iv) Fourth, the raw dataset and 2 component PCA reduced dataset is fitted to various neighboring algorithms like auto, ball tree, kd tree and brute tree of K-Nearest Neighbors classifiers and the performance is analyzed with the confusion matrix.

(v) Fifth, the raw dataset and 2 component PCA reduced dataset is fitted to various kernels of Support Vector Machine classifiers like rbf, poly and sigmoid and the performance is analyzed with the confusion matrix. Firstly, creating the correlation matrix and identifying the relationship between each features in the wine data set.

A. System Architecture

The overall block diagram of the proposed work is shown in Fig. 1.

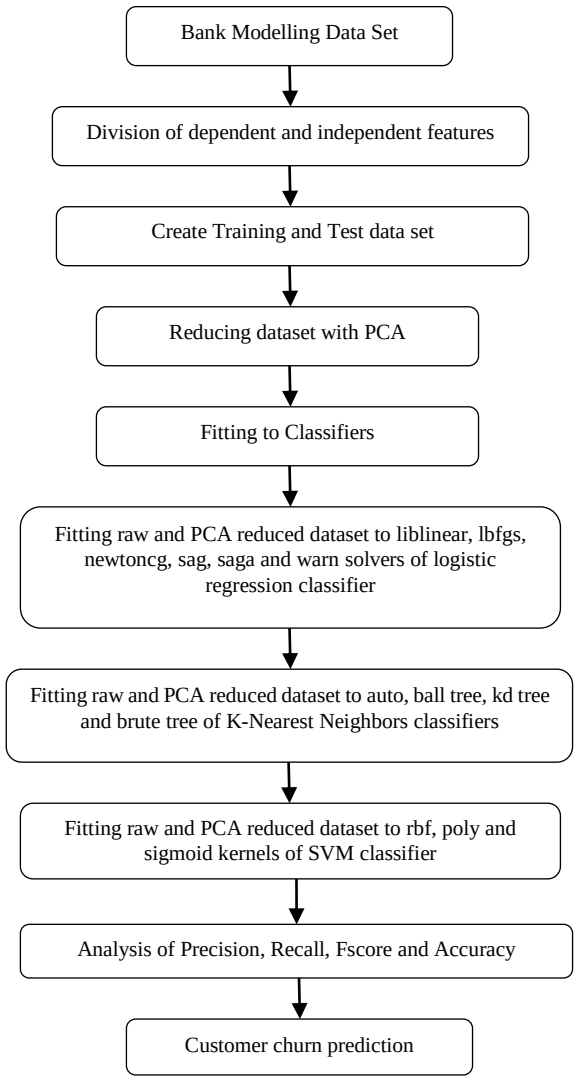

Fig. 1 Overall Design of Work

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Customer Churn Analysis

The customer churn bank modelling data set from UCL ML Repository is utilized for implementation with 8 independent attribute and 1 Exited dependent attribute. The attributes are shown below.

1. CreditScore
2. Age
3. Tenure
4. Balance
5. Number of Products
6. Has Credit Card
7. IsActiveMember
8. Estimated Salary
9. Exited (Yes / No) - Dependent Attribute

The relationship between the variables of the customer churn bank modelling data set is shown in Fig. 2.

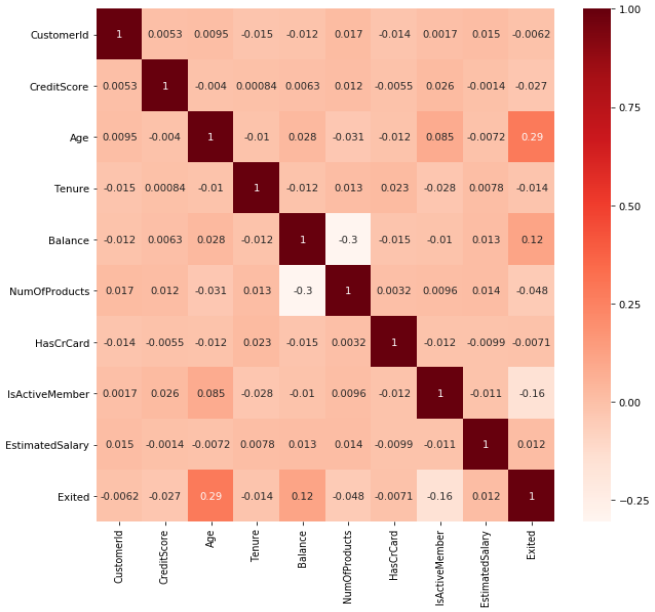

Fig. 2 Correlation Matrix of churn bank dataset

The distribution of the customer churn target variable is shown in Fig. 3.

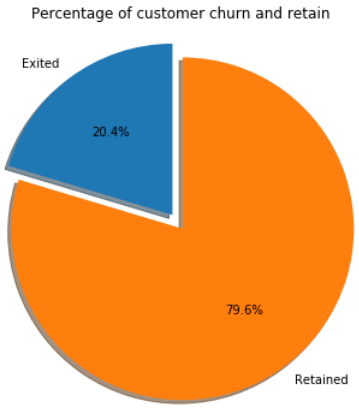

Fig. 3. Customer churn distribution

The relationship of the attributes of the customer churn bank modelling data set with the churn target variable is shown in the Fig. 4 – Fig 10.

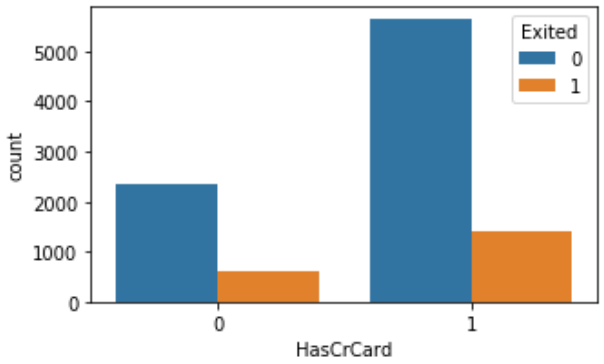

Fig. 4. Has Credit Card VS Customer Churn

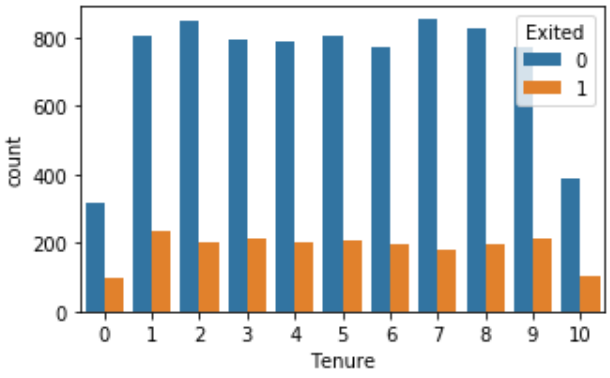

Fig. 5. Tenure VS Customer Churn

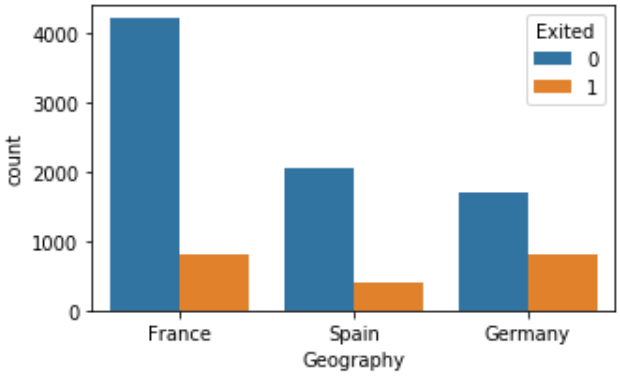

Fig. 6. Has Credit Card VS Customer Churn

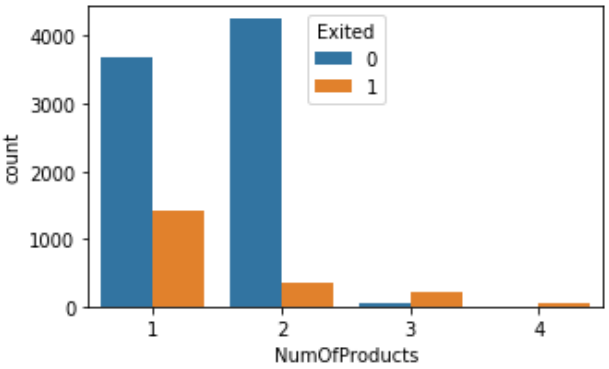

Fig. 7. Number of Products VS Customer Churn

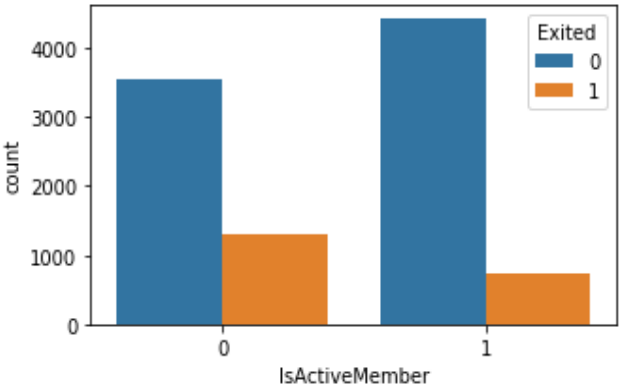

Fig. 8. Is Active Member VS Customer Churn

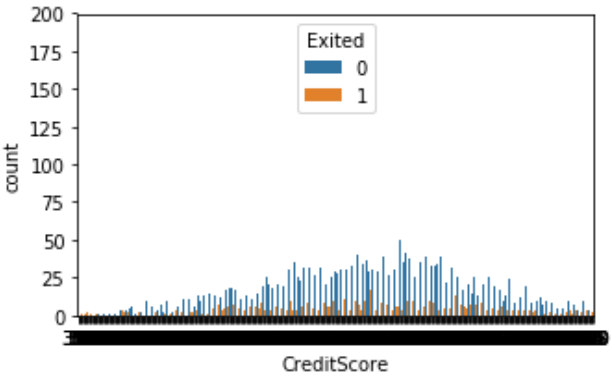

Fig. 9. Credit Score VS Customer Churn

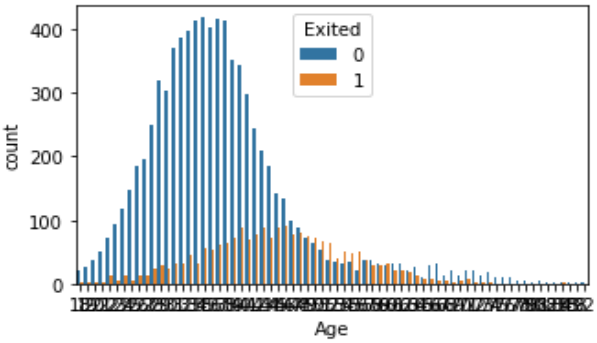

Fig. 10. Age VS Customer Churn

The raw dataset is fitted to various solvers of logistic regression classifiers like liblinear, lbfgs, newtoncg, sag, saga and warn and their performance is analyzed with the confusion matrix and is shown in Fig. 11.

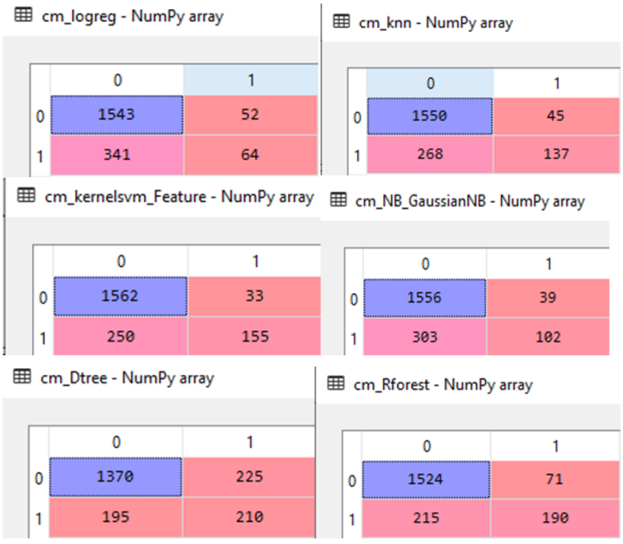

Fig. 11. Various Solvers of Logistic Regression Classifiers

The 2 component PCA reduced dataset is fitted to various solvers of logistic regression classifiers like liblinear, lbfgs, newtoncg, sag, saga and warn and their performance is analyzed with the confusion matrix and is shown in Fig. 12.

Fig. 12 Various Solvers of Logistic Regression Classifiers with PCA

The raw dataset is applied for dimensionality reduction with principal component analysis and the obtained relation wih PCA component is shown in Fig.13

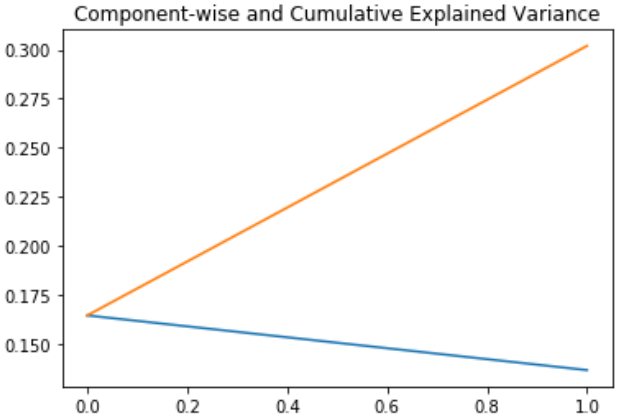

Fig. 13 PCA Components VS Variance

The raw dataset is fitted to various neighboring algorithms like auto, ball tree, kd tree and brute tree of K-Nearest Neighbors classifiers and their performance is analyzed with the confusion matrix and is shown in Fig. 14.

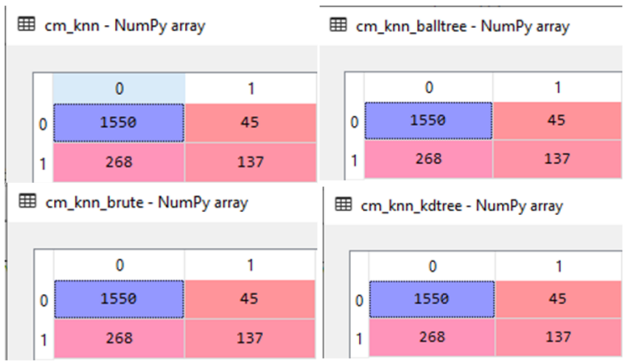

Fig. 14. Various neighboring algorithms of KNN Classifiers

The 2 component PCA reduced dataset is fitted to various neighboring algorithms like auto, ball tree, kd tree and brute tree of K-Nearest Neighbors classifiers and their performance is analyzed with the confusion matrix and is shown in Fig. 15.

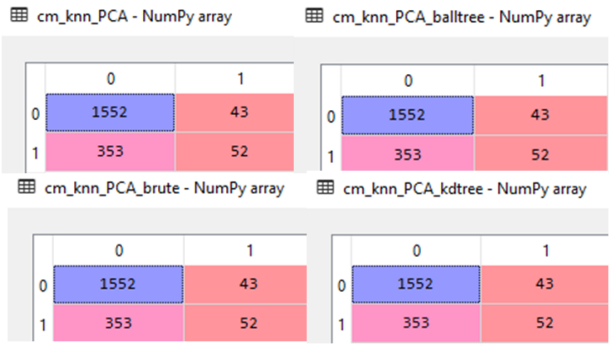

Fig. 15. Various neighboring algorithms of KNN with PCA

The raw dataset is fitted to various kernels like rbf, poly, sigmoid of support vector machine classifiers and their performance is analyzed with the confusion matrix and is shown in Fig. 16.

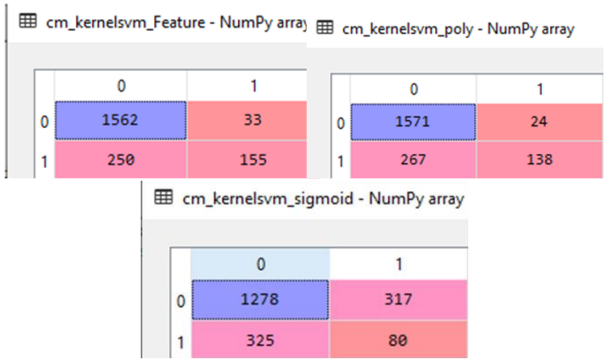

Fig. 16. Various Kernels of SVM classifier

The 2 component PCA reduced dataset is fitted to various kernels like rbf, poly, sigmoid of support vector machine classifiers and their performance is analyzed with the confusion matrix and is shown in Fig. 17.

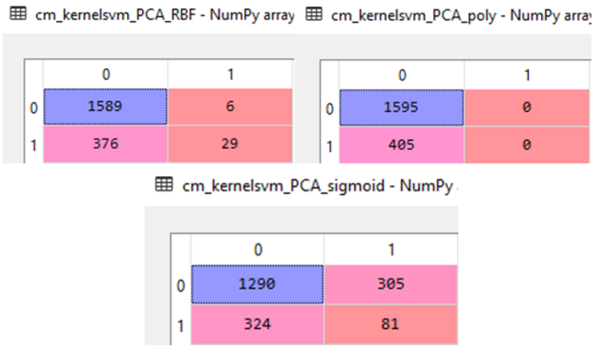

Fig. 17. Various Kernels of SVM classifier with PCA

The performance analysis of the raw and 2 component PCA reduced dataset for Solvers of Logistic Regression classifiers is shown in Table. 1 – Table 3.

Table. 1 Performance Analysis of Solvers of Logistic Regression classifiers without PCA

Solvers	Logistic Regression classifiers		
	Precision	Recall	FScore
Liblinear	0.76	0.80	0.76
Lbfgs	0.76	0.80	0.76
Newtoncg	0.76	0.80	0.76
Sag	0.76	0.80	0.76
Saga	0.76	0.80	0.76
Warn	0.76	0.80	0.76

Table. 2 Performance Analysis of Solvers of Logistic Regression classifiers with PCA

Solvers	Logistic Regression classifiers		
	Precision	Recall	FScore
Liblinear	0.64	0.80	0.71
Lbfgs	0.64	0.80	0.71
Newtoncg	0.64	0.80	0.71
Sag	0.64	0.80	0.71
Saga	0.64	0.80	0.71
Warn	0.64	0.80	0.71

Table. 3 Accuracy Analysis of Solvers of Logistic Regression classifier

Solvers	Logistic Regression classifier	
	Accuracy % without PCA	Accuracy % with PCA
Liblinear	80.3	79.7
Lbfgs	80.3	79.7
Newtoncg	80.3	79.7
Sag	80.3	79.7
Saga	80.3	79.7
Warn	80.3	79.7

The performance analysis of the raw and 2 component PCA reduced dataset for neighbor algorithm of KNN classifiers is shown in Table. 4 – Table 6.

Table. 4 Performance Analysis of neighbor algorithm of KNN classifiers without PCA

Neighbor Algorithm	KNN classifier		
	Precision	Recall	FScore
Auto	0.83	0.84	0.82
Ball Tree	0.83	0.84	0.82
Brute	0.83	0.84	0.82
KDTree	0.83	0.84	0.82

Table. 5 Performance Analysis of Solvers of KNN classifiers with PCA

Neighbor Algorithm	KNN classifier		
	Precision	Recall	FScore
Auto	0.76	0.80	0.75
Ball Tree	0.76	0.80	0.75
Brute	0.76	0.80	0.75
KDTree	0.76	0.80	0.75

Table. 6 Accuracy Analysis of Solvers of KNN classifier

Neighbor Algorithm	KNN classifier	
	Accuracy % without PCA	Accuracy % with PCA
Auto	84.3	80.2
Ball Tree	84.3	80.2
Brute	84.3	80.2
KDTree	84.3	80.2

The performance analysis of the raw and 2 component PCA reduced dataset for kernels of SVM classifiers is shown in Table. 7 – Table 9.

Table. 7 Performance Analysis of kernels of SVM classifiers without PCA

Kernels	SVM classifier		
	Precision	Recall	FScore
Rbf	0.85	0.86	0.84
Poly	0.85	0.85	0.83
Sigmoid	0.68	0.68	0.68

Table. 8 Performance Analysis of kernels of SVM classifiers with PCA

Kernels	SVM classifier		
	Precision	Recall	FScore
Rbf	0.81	0.81	0.74
Poly	0.64	0.80	0.71
Sigmoid	0.68	0.69	0.68

Table. 9 Accuracy Analysis of kernels of SVM classifier

Kernels	SVM classifier	
	Accuracy % without PCA	Accuracy % with PCA
Rbf	85.8	80.9
Poly	85.4	79.7
Sigmoid	67.9	68.55

V. CONCLUSION

This paper attempts to predict the existence of the early customers of the banking sector using various solvers, neighbouring algorithms and kernels of logistic regression, KNN and Support vector machine classifier respectively. The performance of all the classifiers of the wine data set is compared before and after applying PCA and LDA. Experimental results shows that, the rbf kernel of Support vector machine classifier is effective with the accuracy of 85.8% before applying PCA and accuracy of 80.9% after applying PCA compared to other classifiers.

REFERENCES

1. Petkovski, Aleksandar & Risteska Stojkoska, Biljana & Trivodaliev, Kire & Kalajdziski, Slobodan. "Analysis of churn prediction: A case study on telecommunication services in Macedonia". 1-4. 10.1109/TELFOR.2016.7818903, 2016.
2. Jadhav, Rahul & Pawar, Usharani. "Churn Prediction in Telecommunication Using Data Mining Technology", International Journal of Advanced Computer Sciences and Applications. 2. 10.14569/IJACSA.2011.020204, 2011.
3. Hung, Shin-Yuan & Yen, David & Wang, Hsiu-Yu."Applying data mining to telecom chum management. Expert Systems with Applications" . 31. 515-524. 10.1016/j.eswa.2005.09.080, 2006.
4. Pereira, Fabiola & Carvalho, André & Assis, Rafael & Costa, Maxley & Faria, Elaine & Julia, Rita & Barcelos, Umberto & Melo, Jony, " Feature-Based Time Series Classification for Service Request Opening Prediction in the Telecom Industry". 10.1007/978-3-030-30244-3_11, 2019.
5. Adnan Amin, Feras Al Obedat, Babar shah, Awais Adnan, Jonathan Loo, Sajid Anwar, " Customer Churn prediction on telecommunication industry using data certainty", Journal of Business Research, vol. 94, Jan 2019, pp. 290 -301.
6. M. Shyamala Devi, Shakila Basheer, Rincy Merlin Mathew, "Exploration of Multiple Linear Regression with Ensembling Schemes for Roof Fall Assessment using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019.
7. Shakila Basheer, Rincy Merlin Mathew, M. Shyamala Devi, "Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 127-133.
8. Rincy Merlin Mathew, M. Shyamala Devi, Shakila Basheer," Exploration of Neighbor Kernels and Feature Estimators for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 597-605.
9. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Nariboyena Vijaya Sai Ram, "Backward Eliminated Formulation of Fire Area Coverage using Machine Learning Regression", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp.1565-1569
10. M. Shyamala Devi, Ankita Shil, Prakhar Katyayan, Tanmay Surana, "Constituent Depletion and Divination of Hypothyroid Prevalance using Machine Learning Classification", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, October 2019, pp. 1607-1612
11. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Sairam Kondapalli, "Recognition of Forest Fire Spruce Type Tagging using Machine Learning Classification", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, pp. 4309 – 4313, 16 September 2019.
12. M. Shyamala Devi, Usha Vudatha, Sukriti Mukherjee, Bhavya Reddy Donthiri, S B Adhiyan, Nallareddy Jishnu, " Linear Attribute Projection and Performance Assessment for Signifying the Absenteeism at Work using Machine Learning", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, pp. 1262 – 1267, 16 September 2019.
13. M. Shyamala Devi, Mothe Sunil Goud, G. Sai Teja, MallyPally Sai Bharath, "Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.11, pp. 604 – 609, 30 September 2019.
14. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, " Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 952 – 956, 30 August 2019.
15. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 957 – 961, 30 August 2019.
16. Rincy Merlin Mathew, R. Suguna, M. Shyamala Devi, "Composite Model Fabrication of Classification with Transformed Target Regressor for Customer Segmentation using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 962 – 966, 30 August 2019.
17. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 4800-4807.
18. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 6198-6203.
19. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, , vol. 22, no. 4, 25 June 2019, pp. 729-739. DOI: 10.1080/09720510.2019.1609729ISSN: 0972-0510 (Print), 2169-0014 (Online).
20. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739. DOI: 10.1080/09720510.2019.1609729ISSN: 0972-0510 (Print), 2169-0014 (Online).
21. Shyamala Devi Munisamy, Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification", Learning and Analytics in Intelligent Systems, LAIS, Springer, vol. 3, pp. 604-612, June 2019.
22. Suguna Ramadass, Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers", Springer's book series entitled "Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 613-620, June 2019.
23. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
24. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.