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

Postulation of Customer Retention in Banking Sector using Machine Learning and Principal Component

M. Shyamala Devi, G. Bhargava Krishna, K.Sowmya, T. Sabari Pavan

Abstract*: Recently, there is a rapid growth in technological improvement in banking sector. The entire world is using the banking service for managing their financial and property assets. As of now, all the technological advancements are applied to banking sector to facilitate the customers with proper operational excellence. In this view, the bank has complete responsibility in serving the people with their modern application to save their time and wealth. So the customer value analysis is needed for the bank to enrich the marketing growth and turnover of the bank. But still, the prediction of customer churn remains a challenging issue for the banking sector for analyzing the profit growth. With this view, we focus on predicting the customer churn for the banking application. This paper uses the churn 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 set is applied to various classifiers like Logistic Regression, KNN, Kernel SVM, Naive Bayes, Decision Tree, Random Forest to analyze the confusion matrix. The Performance analysis is done by comparing the metrics like Precision, Recall, FScore and Accuracy. Second the data set is subjected to dimensionality reduction method using Principal component Analysis and then fitted to the above mentioned classifiers and their performance analysis is done. Third, the performance analysis is done for the dataset by comparing the metrics with and without applying the dimensionality reduction. A Performance analysis is done with various classification algorithms and comparative study is done with the performance metric such as accuracy, precision, recall, and f-score. The implementation is carried out with python code using Anaconda Navigator. Experimental results shows that before applying dimensionality reduction PCA, the Random Forest classifier is found to be effective with the accuracy of 86%, Precision of 0.85, Recall of 0.86 and FScore of 0.84. Experimental results shows that after applying dimensionality reduction, the 2 component PCA with the kernel SVM classifier is found to be effective with the accuracy of 81%, Precision of 0.81, Recall of 0.81 and FScore of 0.74. 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.
G. Bhargava Krishna, Finar Year B. Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
K.Sowmya, Finar Year B. Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
T. Sabari Pavan, Finar 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

Customer satisfaction analysis and their feedback is directly associated with the financial profit of the company. So the analysis of the customer satisfaction and their feedbach is most essential in the forecasting of the profit and turnover of the company. In the current world scenario, the banking sector have drastically improved and utilized by the entire people to safeguard their money. The Nanking sector are also diverged their contribution in various aspects and everything are computerized. The customers can easily view their transactions online once completed with their work. The customers can update their daily transactions details by accessing their account number. Though, banking sector are providing their maximum service to the people, the expectations of the people always changes. By the new technological development, the customers also expect a lot of new services from the banking sector. This raises the real challenge in the hands of the banking sector to retain their customers with their bank. This leads to analysis of the customer satisfaction and the method of predicting the customer detention for the fore coming years to be a challenging task for the bank organizations. This is the place where the machine learning is used by the banking organizations for analyzing the customer feedback and to predict the customer churn.

The paper is organized in such a way that Section 2 discusses the existing systems for predicting the customer churn analysis. Section 3 reveals 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 group segment prediction and analysis is done for the wine orgnanization by using the methods like Linear Regression, Decision Trees and Artificial Neural Networks [1].

The customer churn analysis for the banking application is performed using Spark ML package by dealing with big data.

The prediction of customer churn is done based on their transaction data [2].

The customer churn analysis can be related to customer relationship management where the bank customers are converted into competitors. This in turn builds the effective customer relationship with the bank management. The prediction of customer churn is done with the logistic regression methods and the non churn customers are analyzed for retail banking organization [3]. The new performance metric was analyzed for the prediction of customer churn for the banking data using machine learning classification [4]. The ensembling methods were also incorporated to analyze the customer churn predictions [5]. The customer churn analyses were also done based on the review data from the social network. The prediction is done with soft computing methods. The customer churn analysis and prediction is implemented with profit optimizing approach using Bayesian network classifiers and data mining algorithms.

III. PROPOSED WORK

In our proposed work, the customer churn bank modelling data set is used for predicting the customer churn. Our implementation in this paper is shown below.

(i) Firstly, creating the correlation between the variables are identified and displayed.
(ii) Secondly, high importance features of the customer churn bank modelling dataset is identified by Adaboost regressor
(iii) Thirdly, analysing the proportions of the customers churn distribution of the target variable.
(iv) Fourth, the customer churn bank modelling data set is fitted to KNN, Random forest, Decision Tree, Kernel AVM, Naive Bayes and Logistic regression classifiers.
(v) Fifth, the variable reduction is done using PCA with number of components as 2. The dimensionality reduced dataset is fitted to KNN, Random forest, Decision Tree, Kernel AVM, Naive Bayes and Logistic regression classifiers.
(vi) Sixth, the performance analysis is compared for step 4 and 5 using the metrics like FScore, Precision, Accuracy and Recall

A. System Architecture

The overall design f our work is shown in Fig. 1

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

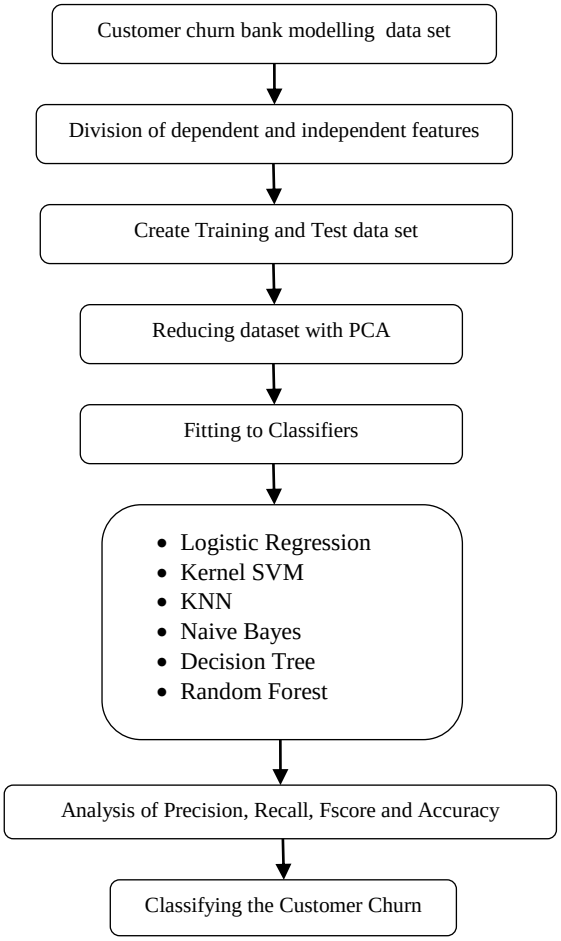

Fig. 1 Overall Design Architecture

Customer churn bank modelling data set is subjected to analyze the relationship between the variable and the distribution is shown as correlation matrix and is shown in Fig. 2.

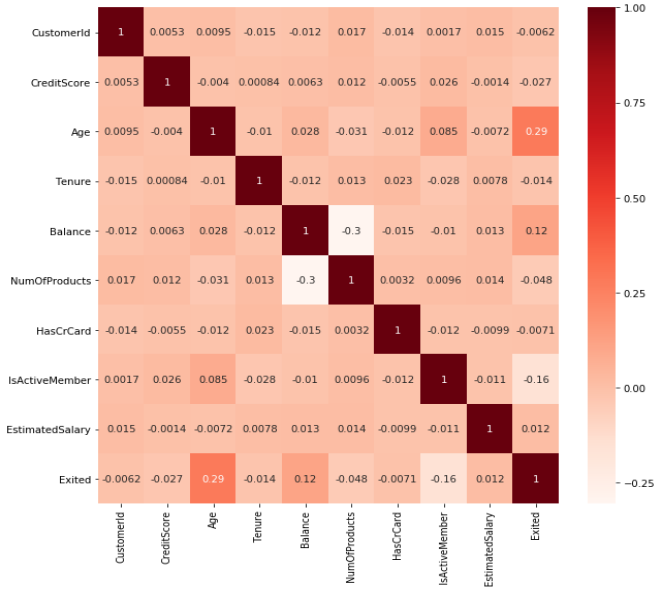

Fig. 2 Correlation Matrix of Bank data set

The important variables of the customer churn bank modelling data set are projected as shown in Fig. 3.

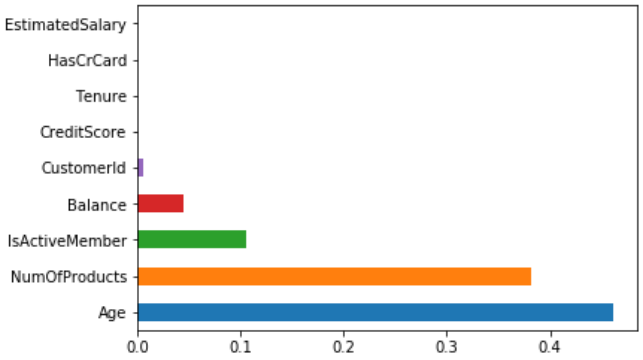

Fig. 3 Important Features of the data set

The probability distribution of important features of the data set is shown in Fig. 4.

AdaBoost_featureimportances - Seri

Index	0
CustomerId	0
CreditScore	0
Age	0.450393
Tenure	0
Balance	0.0604097
NumOfProducts	0.386757
HasCrCard	0
IsActiveMemb…	0.102441
EstimatedSal…	0

Fig. 4 Probability distribution of the dataset attributes

The customer churn target distribution of the customer churn bank modelling data set is shown in Fig. 5.

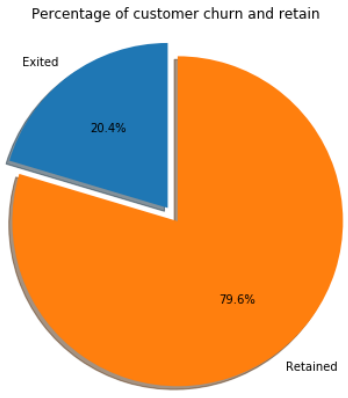

Fig. 5 Customer Churn Target Distribution

The customer churn bank modelling data set is fitted to KNN, Random forest, Decision Tree, Kernel AVM, Naive Bayes and Logistic regression classifiers and the confusion matrix is shown in Fig. 6.

cm_logreg - NumPy array

	0	1
0	1543	52
1	341	64

cm_knn - NumPy array

	0	1
0	1550	45
1	268	137

cm_kernelsvm_Feature - NumPy array

	0	1
0	1562	33
1	250	155

cm_NB_GaussianNB - NumPy array

	0	1
0	1556	39
1	303	102

cm_Dtree - NumPy array

	0	1
0	1370	225
1	195	210

cm_Rforest - NumPy array

	0	1
0	1524	71
1	215	190

Fig. 6 Confusion Matrix for the classifiers

The variable reduction is done using PCA with number of components as 2 and is shown in Fig. 7.

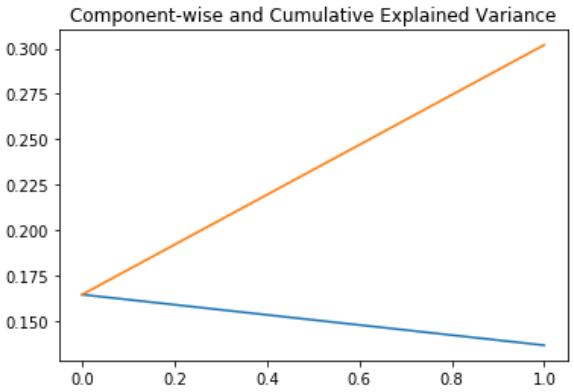

Fig. 7 PCA Component Distribution

The dimensionality reduced dataset is fitted to KNN, Random forest, Decision Tree, Kernel AVM, Naive Bayes and Logistic regression classifiers and the confusion matrix is shown in Fig. 8.

cm_logreg_PCA_liblinear - NumPy array

	0	1
0	1595	0
1	405	0

cm_knn_PCA - NumPy array

	0	1
0	1552	43
1	353	52

cm_kernelsvm_PCA_RBF - NumPy array

	0	1
0	1589	6
1	376	29

cm_NB_PCA_GaussianNB - NumPy array

	0	1
0	1595	0
1	405	0

cm_Dtree_PCA - NumPy array

	0	1
0	1293	302
1	276	129

cm_PCA_Rforest - NumPy array

	0	1
0	1484	111
1	317	88

Fig. 8 PCA Confusion Matrix for the classifiers

The performance analysis is done with metric comparison is shown in the Table. 1, Table. 2 and Table. 3.

Table. 1 Performance Comparison of Precision, Recall and FScore for all the classifiers without PCA

Classifier	Performance Metrics without PCA		
	Precision	Recall	FScore
Logistic Reg	0.76	0.80	0.76
KNN	0.83	0.84	0.82
Kernel SVM	0.85	0.86	0.84
Random Forest	0.85	0.86	0.84
Naïve Bayes	0.81	0.83	0.79
Decision Tree	0.80	0.79	0.79

Table. 2 Performance Comparison of Precision, Recall and FScore for all the classifiers after applying PCA

Classifier	Performance Metrics with PCA for 3 components		
	Precision	Recall	FScore
Logistic Reg	0.64	0.80	0.71
KNN	0.76	0.80	0.75
Kernel SVM	0.81	0.81	0.74
Random Forest	0.75	0.79	0.76
Naive Bayes	0.64	0.80	0.71
Decision Tree	0.72	0.71	0.71

Table. 4 Performance Comparison of Logarithmic Loss and Accuracy for all the classifiers before applying PCA

Classifier	Accuracy % without PCA	Accuracy % with PCA
Logistic Reg	80	79
KNN	84	80
Kernel SVM	85	81
Random Forest	86	79
Naive Bayes	83	80
Decision Tree	79	71

V. CONCLUSION

This paper attempts to predict the customer churn analysis for the banking sector. The customer churn prediction is done with and without applying the dimensionality reduction to the customer churn bank modelling dataset. The dimensionality reduction is done with principal component analysis. Experimental results shows that before applying dimensionality reduction PCA, the Random Forest classifier is found to be effective with the accuracy of 86%, Precision of 0.85, Recall of 0.86 and FScore of 0.84. Experimental results shows that after applying dimensionality reduction, the 2 component PCA with the kernel SVM classifier is found to be effective with the accuracy of 81%, Precision of 0.81, Recall of 0.81 and FScore of 0.74. compared to other classifiers.

REFERENCES

1. Jorge Ribeiro , Jose Neves , Juan Sanchez , Manuel Delgado , Jose Machado, and Paulo Novais, "Wine Vinification prediction using Data Mining tools", Computing and Computational Intelligence, 2016
2. Hend Sayed, Manal A. Abdel-Fattah, Sherif Kholief," Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study", International Journal of Advanced Computer Science and Applications, Vol. 9, No. 11, 2018 Information Systems Department Faculty of Computers and Information Helwan University Cairo, Egypt
3. Verbraken, T., Verbeke, W., Baesens, B., "A novel profit maximizing metric for measuring classification performance of customer churn prediction models".IEEE Transactions on Knowledge and Data Engineering 25 (5), 2013, pp. 961–973.
4. Van Wezel, M., Potharst, R., "Improved customer choice predictions using ensemble methods". European Journal of Operational Research 181 (1), 2007, pp. 436–452.
5. Verbeke W, Martens D, Baesens B, "Social network analysis for customer churn prediction", Applied Soft Computing 14, part C:431446, 2014
6. Verbraken T, Verbeke W, Baesens B, "Profit optimizing customer churn prediction with Bayesian network classifiers". Intelligent Data Analysis 18(1), 2014, pp. 3–24.
7. 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.
8. 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.
9. 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.
10. 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
11. 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
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.

18. 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.
19. 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.
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. 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).
22. 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.
23. 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.
24. 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.
25. 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.