ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3955–3962
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.76 Spec Issiue on ICAML-2020

A SURVEY OF EMPLOYEE AND CUSTOMER CHURN PREDICTION METHODOLOGIES

S. MADANE AND D. CHITRE[1]

ABSTRACT. Nowadays the issue of employee or customer churn has become a crucial one for various companies. One of the major concerns of the companies is to diminish employee attrition rates since the cost of replacing previous employees is high. Also, customer churn needs to be minimized to save the companies from huge financial losses. One way oh handling customer/employee churn is the prediction of the tendency of various customers/employees to leave a particular organization. Various classification models in data mining can be used for churn prediction. This research focuses on surveying the various such prediction methods like decision trees, logistic regression, Support Vector Machines (SVM), K-Nearest Neighbour (KNN), random forest, Naive Bayes, Artificial Neural Network (ANN), penalized multi-criteria linear programming (PMCLP) and feature selection based transfer ensemble model (FSTE) and comparing the performance of these nine methods in terms of their average accuracy. It is found that random forest method gives the highest average accuracy of 93.16% and hence performs the best.

1. INTRODUCTION

Employee or customer churn is a serious problem in organizations. Customer churn refers to the tendency of customers to stop the usage of the products or

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 90B50, 62C05.

Key words and phrases. employee churn, customer churn, prediction, classification, churn rate, attrition, Support Vector Machine, decision tree, random forest, Naive Bayes, feature selection.

services of a particular company, [1]. This can be causes by a customer's dissatisfaction with the quality of product/service or the failure of the product/service to meet the customer's expectations. This results is the customers switching to the products and services offered by other companies. Employee churn or employee attrition refers to the turnover in a company, [1]. It is the result of the current employees leaving the company due to job dissatisfaction. An employee can experience job dissatisfaction due to several reasons like not getting promotion, no increment in salary for a substantial period of time, excessive workload etc. [2, 3]. Also, employees may switch their jobs due to inconvenient job location, unhealthy atmosphere at workplace, insufficient holidays and job incentives etc. To identify such patterns and reasons, histograms and correlation matrices can be used [3]. IT organizations face an employee churn rate of 12-15%, [4].

Employee churn causes a serious problem for the companies since when employees leave the company, it becomes essential to recruit new employees to replace the previous ones, [1]. Also, the task of recruiting new employees and training them is expensive and requires the efforts and attention of the company officials. Thus, the company faces a financial loss and consequently fails to provide sufficient salary and incentives to its employees, which can result in job dissatisfaction among the employees and a resulting turnover in the company.

Several techniques are used by the organizations to minimize their employee/customer attrition rate by performing churn prediction. A technique that is widely used is data mining, [1]. There are several classification models like Support Vector Machines (SVM), linear and logistic regression, decision trees, random forest, [5], Classification and Regression Trees (CART), [6], K-Nearest Neighbour (KNN), Naive Bayes classifier, multi-criteria linear programming (MCLP), [7], Neural Network, [8], Fuzzy Logic, evolutionary algorithms, transfer learning, [9] etc. can be used for prediction based on classification. Also, methods like automated Business Process Discovery, [10], can be used for churn prediction.

This aim of this paper is to survey the various such methodologies for churn prediction. There are nine methods that we have surveyed and compared in terms of performance. The section 2 provides a detailed literature survey of various classification techniques in data mining that are used for customer/employee

churn prediction. In section 3, we have compared the papers considered for conducting this survey in a tabular form, in terms of the methodology, dataset used, accuracy, research results and limitations or future scope. Further, in section 4 we have analyzed our survey and compared the various methods covered in the survey in terms of their average accuracy. In section 5, we have concluded our research by drawing useful inferences.

2. Literature Survey

Ibrahim Onuralp et al. consider various data mining techniques and procedures to predict the employee churn pertaining only to employees who leave voluntary. Support vector machine proves to be the best method in terms of accuracy and precision whereas the Decision Tree approach gives a higher recall. The paper also talks about feature selection to consider only the most relevant features from the dataset obtained from IBM having 1470 records and 34 features, [2].

Andry Alamsyah et al. consider the data obtained from a telecommunication company in Indonesia. They use a 4 stage process for the prediction of employee churn starting from data collection from Human Resource Information System which consisted of 12 attributes. A classification model was built using the Naive Bayes, Decision Tree and Random Forest. A confusion matrix was prepared for each methodology with Naive Bayes showing the highest number of true positives. However, Random Forest had the highest accuracy of 97.5% and was declared as the best prediction method, [1].

Sepideh Hassankhani Dolatabadi et al. discuss an approach for designing a customer and employee churn prediction model using data mining and Neural Network methodologies. A dataset over a period of 1.5 years was considered having 21 attributes. A total of 9239 records were taken into account out of which 24.2% were churned. 15 attributes were considered for customer prediction. Decision Tree, Naive Bayes, SVM and Neural Network methods were applied for the prediction. Naive Bayes had the fastest information processing speed but SVM had the highest accuracy and the True Positive rate followed by the Bayesian approach, [8].

In [4], V. Vijaya Saradhi et al. perform employee churn prediction using three classification algorithms: SVM, Naive Bayes and random forest. The dataset includes the details of all employees in a particular client unit of an organization.

The dataset contains 25 attributes and there are three possible output classes: resigned, released and retained. Irrelevant attributes were eliminated and derived attributes were used. The results showed that random forest gave the higher accuracies than SVM and Naive Bayes. But, the SVM defeated the other two methods by giving high true positive rates and this was because penalties were assigned to individual classes to make SVM tolerant to the class imbalance problem in the dataset.

The customers with low credit scores were classified as churners. The lift curve was plotted for various algorithms and a comparison of the curve for the different random forest methodologies using which it can be seen that IBRF performs significantly better in terms of speed, training rate and scalability, [5].

In [7], Aihua Li et al. discuss a method of performing churn analysis on a website/company email users by using two methods, namely, PMCLP (penalized multi-criteria linear programming), and decision tree. Using these methods, the loyal customers can be distinguished from the ones who tend to leave. For each class (good and bad), ideal values are determined, and then regret measures are calculated based on the deviation of the actual values of the samples from the ideal values.

In [6], Chuanqi Wang et al. propose a cost-sensitive CART model to reduce the costs due to misclassification. A Cost Partition Attribute(CPA) is determined based on which the dataset is partitioned into blocks, and then for each block, the cost matrix based on CPA and a cost sensitive CART model (CARTCS) are developed.

In [10], Edward Peters et al. have performed a case study regarding the identification of causes of customer churn in the 'care-type' services. According to [10], customer churn results from degradation in two main factors, the quality of services and the speed of services. Automated Business Process Discovery was performed using Comprehend dataset.

3. Discussion

Table 1 provides a comparison of the various methodologies adopted in the existing approaches. The parameters of comparison are the dataset used, accuracy obtained, results, and limitations of the particular approach.

Table 1: Discussion

METHODOLOGY & DATASET	ACCURACY	RESEARCH RESULTS	LIMITATIONS AND GAPS
Used Decision Tree, Logistic Regression, SVM, KNN, Random Forest and Naive Bayes with & without feature selection [2] **Dataset:** International Business Machines Corporation	Logistic Regression without feature selection had highest accuracy of 87.1%, SVM feature selection had an accuracy of 89.7%	SVM had the highest accuracy and precision with values 89.7% and 51%, whereas Decision Tree had the highest recall of 43%.	Churn probability for each employee can be calculated and given score for retention
Naive Bayes, Decision Tree, RFC methodologies were used to build a model for comparison of the best method for employee churn prediction [1] **Dataset:** Indonesian telecommunication company database	RFC showed the highest accuracy of 97.5%.	Random Forest is the best method for prediction followed by Naive Bayes and Decision Tree on the basis of accuracy, F1 score, precision and recall values.	Not addressed
Decision Tree, Naive Bayes, SVM, Neural Network methodologies used for employee and customer prediction [8] **Dataset:** Employee and customer database	SVM showed 99.83% accuracy, highest for customer churn values, all methods except SVM show 100% accuracy for customer churn.	100% accuracy was obtained for prediction of employee churn, Processing speed and Analysis power was considered for model evaluation.	Thorough inspection of customer and employee features is needed to improve the prediction models

SVM, RF and Naive Bayes are used for employee churn prediction. Customer Lifetime Value (CLV) and Employee Value Model (EVM) are used to distinguish between churners [4]. **Dataset:** Client unit employees database	Highest accuracy is obtained by RF (83.49%, 92.28% and 88.81% for Models I, II and III respectively).	Random forest achieves the highest accuracy and True Negative Rate whereas SVM performs best in terms of True Positive Rate	Improved input data representation can result into higher prediction accuracy, employee work history can be represented in an event oriented manner
Improved Balanced Random Forest method which is a combination of weighted and balanced random forest method was utilized for predicting customer churn, [7]. **Dataset:** Chinese bank database	IBRF has an accuracy of 93.24% which is greater than the accuracies of ANN, Decision tree and SVM.	Lift curves and top decile lifts are used for model evaluation. IBRF has the best top-decile lift of 7.1 and is more scalable and faster than the others.	Not cost effective, effectiveness and generalization ability can be improved

4. Performance Analysis

Thus, the from fig 1 analysis,We come to an conclusion that random forest can be considered as the best classification algorithm for churn prediction. Algorithms like SVM, KNN, Naive Bayes and ANN are also seen to yield decent accuracies. But decision tree gives lower accuracy than most of the other data mining models, and hence it is not an ideal choice for churn prediction.

5. Conclusion

Nowadays customer and employee churn is an crucial issue in organizations since it leads to huge monetary losses. Companies need to reduce the rate of

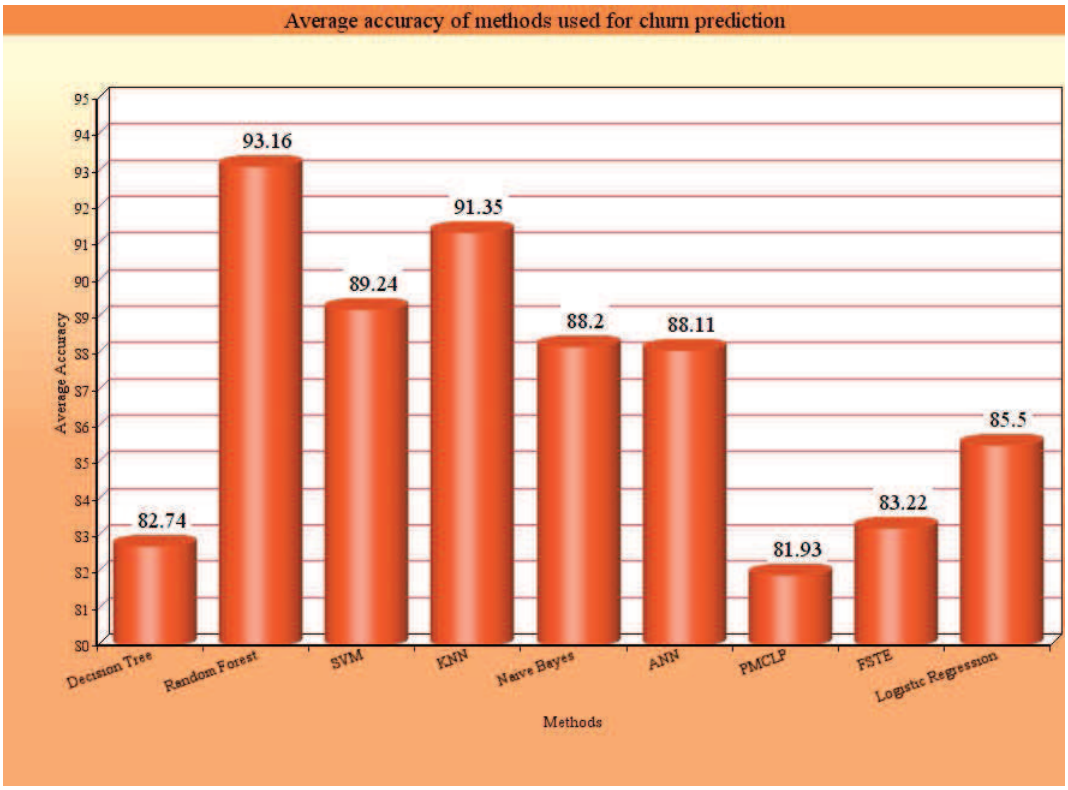

FIGURE 1. Accuracy of Various Methodologies

attrition of their employees and customers and for this, prediction of the tendency of customers and employees to churn is an effective method. Several classification algorithms in data mining can be used for giving reliable and accurate churn prediction. In this research paper we have surveyed such methods and compared their performance in terms of their average accuracy. From the performance analysis, it can be concluded that random forest gives the maximum average prediction accuracy of 93.16%, and hence is the best choice for designing accurate employee or customer churn prediction models. Methodologies like KNN, Naive Bayes, ANN and SVM are found to produce decent average prediction accuracies and hence they can also be used to build reliable churn prediction models. On the other hand, decision tree is seen to yield a poor performance in comparison with almost all other methods, with a comparatively lower accuracy of 82.74%.

REFERENCES

[1] A. ALAMSYAH, N. SALMA: *A comparative study of employee churn prediction model*, 4th International Conference on Science and Technology (ICST), (2018), 1–4. IEEE, 2018.

[2] I. O. YIGIT, H. SHOURABIZADEH: *An approach for predicting employee churn by using data mining*, 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), (2017), 1–4.

[3] D. S. SISODIA, S. VISHWAKARMA, A. PUJAHARI: *Evaluation of machine learning models for employee churn prediction*, International Conference on Inventive Computing and Informatics (ICICI), (2017), 1016–1020.

[4] V. V. SARADHI, G. K. PALSHIKAR: *Employee churn prediction*, Expert Systems with Applications, **38**(3) (2011), 1999–2006.

[5] W. YING, X. LI, Y. XIE, E. JOHNSON: *Preventing customer churn by using random forests modeling*, IEEE International Conference on Information Reuse and Integration, (2008), 429–434.

[6] C. WANG, R. LI, P. WANG, Z. CHEN: *Partition cost-sensitive cart based on customer value for telecom customer churn prediction*, 36th Chinese Control Conference (CCC), (2017), 5680–5684.

[7] A. LI, Z. LIN: *Email users churn analysis based on pmclp and decision tree*, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, **7** (2009), 348–350.

[8] S. H. DOLATABADI, F. KEYNIA: *Designing of customer and employee churn prediction model based on data mining method and neural predictor*, 2nd International Conference on Computer and Communication Systems (ICCCS), (2017), 74–77.

[9] L. XIE, D. LI, J. XIAO: *Feature selection based transfer ensemble model for customer churn prediction*, International Conference on System science, Engineering design and Manufacturing informatization, **2** (2011), 134–137.

[10] E. M. PETERS, G. DEDENE, J. POELMANS: *Understanding service quality and customer churn by process discovery for a multi-national banking contact center*, IEEE 13th International Conference on Data Mining Workshops, (2013), 228–233.

Email address: madanesneha44@gmail.com

Email address: dnyanobachitre@ternaengg.ac.in