Churn Prediction using Neural Network based Individual and Ensemble Models

[1]Mehpara Saghir, [2]Zeenat Bibi, [3]Saba Bashir, [4]Farhan Hassan Khan
[1,2,3]Department of Computer Science,
Federal Urdu University of Arts, Science & Technology, Islamabad, Pakistan
[4]Knowledge & Data Science Research Center, College of Electrical & Mechanical Engineering
National University of Sciences & Technology (NUST), Islamabad, Pakistan
{[1]mehparasaghir, [2]beenishblooch786, [3]saba.bashir3000, [4]mrfarhankhan}@gmail.com
[4]farhan.hassan@ceme.nust.edu.pk

Abstract – Churn prediction is still a challenging problem in telecom industry. Many data mining techniques have been employed to predict customer churn and hence, reduce churn rate. Although a number of algorithms have been proposed, there is still room for performance improvement. Therefore this paper evaluates existing individual and ensemble Neural Network based classifiers and proposes an ensemble classifier which utilizes Bagging with Neural Network in order to improve performance measures resulting in better accuracy for churn prediction. This work employs two benchmark datasets, obtained from GitHub, for comparison and evaluation of the proposed model. An average accuracy of 81% is achieved by the proposed model.

Keywords – Churn Prediction, Ensemble, Majority Vote, Bagging, AdaBoost, Cross Validation.

1. INTRODUCTION

In telecom industry, churn is defined as the number of customers who leave their subscribed organization and migrate to another organization [1]. Customer churn prediction is an extensively used phenomenon to retain its valuable customers as it costs five to seven times less as compared to attaining new customers. Usually, the urge of exploitation of better services leads to customer churn. If customer churn rate can be decreased to 5-10% company's growth rate increases up to 30-85% [2]. Churning and retention of customers is part of customer churn management system. Different researchers have implemented customer churn management in service industry using various techniques such as data mining techniques, machine learning algorithms and statistical models.

Data mining is a process applied on data, having previously unknown information which may be useful for taking some decision. More precisely we can say that it works on the large datasets from databases to extract the relevant patterns which may help to design different strategies for a profitable business. Association rules, Support Vector Machine (SVM), Artificial Neural Network (ANN) and Decision Trees (DT) are commonly used data mining techniques using which researchers have proposed many churn prediction models. Machine Learning is emerged from the study and construction of algorithms which emerged from the studies of computational learning and pattern recognition in artificial intelligence. Machine learning algorithms have been used by many researchers in development of prediction systems [3]. Such algorithms are divided into two main categories namely supervised and unsupervised learning algorithms. Classification and regression algorithms fall in supervised learning algorithms category. Classification models are trained on data from past customers to generate a model which is then applied to classify unseen patterns. On the other hand, clustering algorithms focus on similar features to group data in a cluster and then categorize unseen data into any one of the related cluster [4]. Other studies evaluate different factors such as churn rate, prediction performance and customer retention capabilities [5]. These are the main drivers to predict customer behaviour. Many researchers have proposed various solutions for retaining customers by applying different techniques [6]. Inductive algorithms, hybrid oversampling and under sampling, SMOTE oversampling technique, echo state network (ESN) with SVM training algorithm, clustering techniques, artificial neural networks, decision trees, random forest algorithms, PSO, mRMR, GA and ensemble classifiers have been used for achieving performance in churn prediction [2, 6-8]. Early prediction of customer churn increases the chances to retain the customer.

Research Contributions

The contributions made by this research are summarized here:-

- This research evaluates different versions of Neural Network based individual classifiers for churn prediction. These classifiers include Neural Network (NN), Multi-Layer Perceptron (AutoMLP) and Deep Learning (DL).
- These classifiers are then evaluated in an ensemble setting where well-known ensemble methods are utilized such as Bagging, AdaBoost and Majoirty Voting.
- Results are computed and evaluated using various performance measures such as accuracy, precision, recall, f-measure and statistical measures such as kappa, absolute error, relative error and classification error.

Rest of the paper is organized as follows: Section 2 describes literature review. The proposed research model is presented in Section 3. Section 4 comprises of the results, evaluation and discussion. Finally, section 5 concludes the research work providing future research directions.

2. LITERATURE REVIEW

A number of researchers have presented various techniques for customer churn prediction. This section critically discusses state of the art research. An overview of these techniques is presented in Table 1.

A six step data mining model was introduced in [7] to predict customer churn. This paper also provided a

978-1-5386-7729-2/19/$31.00©2019 IEEE

comparison of some traditional techniques including SVM, DT, Regression Analysis (RA), NN and Fuzzy Logic (FL). SVM outperformed other algorithms for churn prediction. [9] concluded that deep learning algorithm can be used to achieve better results. Idris et al. [10] proposed a model for churn prediction based on Genetic Programming (GP) algorithm and AdaBoost. GP programming worked efficiently for searching data and AdaBoost was used in an iterative approach to identify different factors for customer churn behaviour. Particle Swarm Optimization (PSO) under sampling method was used to balance the dataset. This model worked very efficiently for solving many complex problems. Effective feature selection techniques may be implemented using machine learning or deep learning for improving the performance. Anjum et al. [11] proposed a new e-churn model for the improvement of churn prediction of customer by increase in recall. This model used ensemble technique for churn prediction. Different combination were evaluated for algorithms such as C5, QUEST, CHAID, CRT and logistics regression where best results were obtained by the combination of C5 and QUEST which was 93.4%. No dataset was mentioned for the experimentation and evaluation. The accuracy of these algorithms can be increased if customer data history is incorporated. A comparative study [12] was conducted on 10 different techniques to analyse churn prediction models. Ensemble based techniques (Random Forest and Adaboost) outperformed other algorithms with an accuracy of 96% on the dataset containing 3333 records. SVM and Multi-layer perception (MLP) show second-best performance on given dataset. However, small datasets had been employed for the results computation. Improved accuracy can be obtained by employing hybrid or deep learning models.

Verbek et al. [13] proposed a rule based classification technique which uses the ant miner+ and ALBA for outperforming the churn prediction results. These techniques paid special focus on the predictive accuracy, comprehensibility and justifiability. [14] discussed various causes due to which customers leave a company and try to predict churn customers by defining behavioural attributes of customers. This research utilized data mining techniques, machine learning, pattern recognition, support vector machine, statistics and also applied clustering such as k-means and DB scan algorithms to realize the accurate behaviour of customer. Extensive memory usage for small amount of data is one of the core drawbacks of this research. [6] discussed three main factors i.e. prediction performance, churn rate and retention capability used in profit model. The analysis of potential retention model is presented for profit maximization. The research results show that profit and retention have monotonous relationship for prediction. The better profit can be achieved by enhancing the retention capability. Boosting was used to improve accuracy. To increase customer retention, a series of experiments were presented in [15]. Big ML and Azure ML platforms were used for churn prediction by focusing on user services based on demographic and behavioural data.

SVM and echo state network training algorithms were employed in predicting customer churn using the data of telecom companies [16]. Accuracy, good generalization and ability to resist over-fitting problem helped retaining existing customers. A clustering algorithm presented in [1] included semantic driven subtractive clustering method (SDSCM) that was based on subtractive clustering method (SCM) and axiomatic fuzzy sets (AFS). The new algorithm was helpful to improve the accuracy and decrease the risk level. The proposed SDSCM provided efficient results as compared to k-mean clustering algorithm. In the proposed SDSCM, attributes were selected then neighbourhood radius was determined and finally cluster number was computed.

Although a number of techniques have been presented in recent literature, it is evident from the review provided in this section that each approach has its pros and cons. This research is aimed at improving the performance of customer churn prediction.

TABLE 1: Summarized Literature Review

Ref./Year	Techniques	Dataset	Accuracy /Sensitivity
[13] 2011	Ant Miner+, ALBA	Dataset from KDD library	74%
[9] 2012	SVM, NN, DT	23 att & 5000 instances	83.7% ,77.9%,83.7%,
[6] 2015	DL, LR	BUANBXI mobile 20mm Ltd,	-
[7] 2015	SVM, MLP, NB, C4.5, IDK	Local, KDD	SVM= 98.7%,SVM=91.5%
[10] 2017	GP-AdaBoost, PSO	Orange (50k),Cell2Cell(40k)	83.5%,90.4%
[11] 2017	DSS, C5+QUEST	Raw call detail records	93.40%
[14] 2017	SVM, DT, LR	Local	78.3% &79.9%,80.1%
[16] 2017	LR, NN, DL	AzureML, BigML	DL=70.8%,NN=74.6%
[12] 2018	RF, AdaBoost, NN, SVM	3333 records	96% & 94%

3. PROPOSED METHODOLOGY

An overview of the proposed methodology is visually represented in Figure 1 and explained in detail in the following subsections.

3.1 Data Acquisition

Selection of data is a very tricky step and telecom datasets are not commonly available. We have utilized two datasets available freely for research purposes. Both datasets are obtained from github.[1,2]

3.2 Data pre-processing

Pre-processing techniques are applied on the datasets so that we get best performance results. In pre-processing techniques we handled the missing values, remove duplicates, set role to the give attributes and remove outliers. Before using it for our experiments we pre-process and convert into a balanced dataset.

3.3 Classifier Training and Testing

We have applied 10 fold cross validation technique to generate results. In cross validation we have applied three different classifier named as Deep Learning, Neural Network and AutoMLP. Firstly the model is trained on the training dataset then we apply this model for test dataset in the last step we obtained the performance of our model. Details for each of the classifier is given below.

3.3.1 Deep Learning (DL)

DL is basically a machine learning algorithm subjective to the area of AI. It is basically an inspired version of ANN. This model consists of larger architecture in which multiple nodes are connected to one another and they work in a manner like neuron of the human brain. The main objective of training the classifier is to minimize the number of errors [21]. To reduce the error we applied Zero_one loss for some the unseen instances.

$$\boldsymbol{f}: \boldsymbol{R^D} \rightarrow \{\mathbf{0}, \ldots . \boldsymbol{L}\} \qquad [1]$$

Zero_one loss will the written as

$$\boldsymbol{l_{0,1}} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{|D|}} \boldsymbol{I}_{\boldsymbol{f}\left(\boldsymbol{x^{(i)} \neq y^{(i)}}\right)} \qquad [2]$$

Here D is the training dataset where

$$\mathbf{D} \cap \mathbf{D_{train=\varphi}} \qquad [3]$$

Due to the reservation of its optimization for large nodes it is very expensive so instead of it we mostly prefer to use Negative log-likelihood loss function as

$$\boldsymbol{NLL(\theta, D)} = -\sum_{\boldsymbol{i=0}}^{\boldsymbol{|D|}} \mathbf{log}\, \boldsymbol{P}\left(\boldsymbol{Y} = \frac{\boldsymbol{y^{(i)}}}{\boldsymbol{x^{(i)}}}, \boldsymbol{\theta}\right) \qquad [4]$$

NLL works well for the DL classifier.

3.3.2 Neural Nets

The next step is building model on NN which is also one of the machine learning techniques used commonly. Neural nets work in the same concept as the neuron cells of the human brain work. They make a network in which different adjacent layers are connected together. These layers connectively make a deep network. In NN weight of the signals move across the neurons is collected at the output then results are derived from it. NN can work on both types of training mechanism as supervised and unsupervised.

3.3.3 AutoMLP

AutoMLP is our third machine learning classifier. In this classifier NN can adjust its size and learning rate during the training of classifier. This algorithm use some of the ideas from genetic algorithms and some from stochastic optimization.

3.4 Ensemble Classifier

Classifiers are the individual models use training dataset to train them. Among all classifiers some are classified as ensemble classifier because they work in combination with some other classifier and use the working strategy to combine each of them to get the higher precision, recall and classification performance [23]. In our research work we evaluate three different types of ensemble classifier (Bagging, AdaBoost & Majority Voting) and choose the best one ensemble technique for our churn prediction problem. The conceptual approach is presented in Figure 2.

3.4.1 Bagging

Bootstrap aggregation classifier work for the classification and prediction problem solving. It reduces classification error and improves the accuracy in comparison to individual classifiers [23].

3.4.2 Majority Voting

Majority voting, also known as plurality voting, works on multiple classifier feed, gets the high frequency vote from each classifier and calculates the best voting results for the classifiers [23].

3.4.3 AdaBoost

AdaBoost is basically an iterative algorithm that works with some weak classifiers to solve the classification problems by building a strong classifier rather than regression. AdaBoost improves the performance of machine learning classifiers up to many folds[3]. AdaBoost works on the basis of weights assigned to each instance in the training dataset. Mathematical representation of the formula is

$$\boldsymbol{H(x)} = \boldsymbol{sign}\left(\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{\alpha_t}.\boldsymbol{M_t}\,\boldsymbol{(x)}\right) \qquad [5]$$

[1] https://raw.githubusercontent.com/ZiHG/Customer-churn-prediction/master/Data.csv [Last Accessed: Sep 21, 2018]

[2] https://github.com/DionysiosZelios/Predicting-CustomerChurn/blob/master/Churnsimple_approach.csv [Last Accessed: Sep 21, 2018]

[3] https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/ [Last Accessed: Sep 21, 2018]

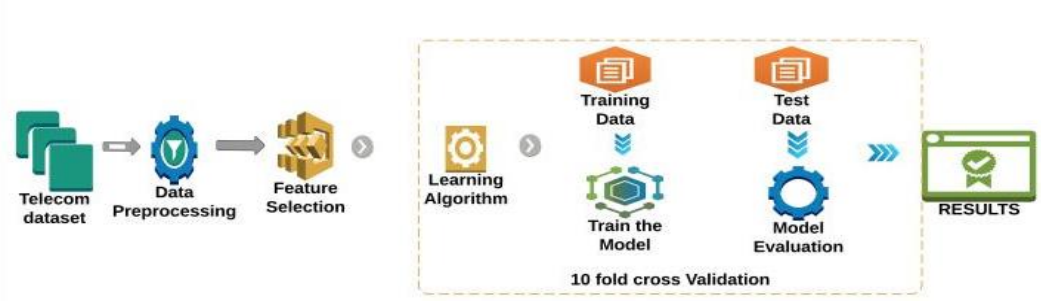

Figure 2: Proposed Model Diagram

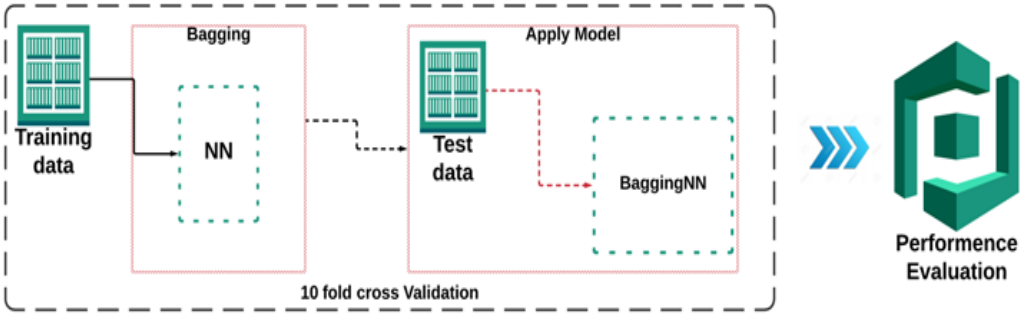

Figure 1: Ensemble based Model Diagram

M_t represent classification of a particular instance based voting for all classifier and α_t is the weight [23].

4. RESULTS, EVALUATION AND DISCUSSION

We have used RapidMiner tool for computation of results. Results were computed on the limited version of RapidMiner which only worked with 10,000 records for the training and evaluating the model. We have applied the classifiers, named above in section 2, for our two datasets one having 3333 records with 23 attributes and the other having 20,000 records along with 12 attributes.

Among all the individual classifiers DL, NN and AutoMLP and ensemble classifier Bagging, AdaBoost and Majority voting with DL, NN and AutoMLP we have observed that Bagging with NN outperforms among all the classifiers in respect to overall average performance of all the performance measures criteria.

The performance results for dataset 1 and 2 are presented in Tables 2 and 3, respectively, whereas Table 4 presents results averaged over the two datasets. By applying the different classification techniques with different classifier on dataset 1 we acquire the following results with DL 91.42%, NN 93.94%, AutoMLP 93.91%, Bagging DL 91.51%, AdaBoost DL 91.09%, Bagging NN 94%, AdaBoost NN 93.07%, Bagging MLP 94.15%, and AdaBoost MLP 93.88%.

Table 2: Performance Measure of Dataset 1

		Accuracy	Precision	Recall	F-Measure	Kappa	Absolute Error	Relative Error	Classification Error
Individual Classifiers	**DL**	91.42	83.035	81.655	82.34	0.646	0.107	10.66	8.58
	NN	93.94	89.955	84.415	87.10	0.736	0.063	6.28	6.06
	AutoMLP	93.91	89.445	84.920	87.12	0.739	0.075	7.52	6.09
Ensemble Classifiers	**Bagging DL**	91.51	90.670	72.940	80.84	0.571	0.110	10.99	8.49
	AdaBoost DL	91.09	82.160	81.550	81.85	0.046	0.114	11.39	8.91
	Bagging NN	**94.00**	**92.475**	**86.200**	**89.23**	**0.778**	**0.067**	**6.72**	**5.07**
	AdaBoost NN	93.07	87.370	83.480	85.38	0.704	0.077	7.75	6.93
	Bagging MLP	**94.15**	**92.230**	**82.995**	**87.37**	**0.733**	**0.080**	**8.04**	**5.85**
	AdaBoost MLP	93.88	89.410	84.815	87.05	0.737	0.081	8.13	6.12
	Majority Voting DL+NN+MLP	93.85	82.285	83.765	83.02	0.728	0.072	7.23	6.15

Table 2: Performance Measure of Dataset 2

		Accuracy	Precision	Recall	F-Measure	Kappa	Absolute Error	Relative Error	Classification Error
Individual Classifiers	**DL**	57.08	61.53	57.09	59.23	0.14	0.5	49.99	43.92
	NN	67.15	70.11	67.15	68.60	0.34	0.37	36.91	32.86
	AutoMLP	67.24	70.6	67.25	68.88	0.35	0.38	37.95	32.26
Ensemble Classifiers	**Bagging DL**	50	25	50	33.33	0	0.5	50	50
	AdaBoost DL	56.72	58.76	65.93	62.14	0.13	0.48	48.29	43.29
	Bagging NN	**67.9**	**70.65**	**68.41**	**69.51**	**0.15**	**0.37**	**37.06**	**32.6**
	AdaBoost NN	66.99	67.61	71.21	69.36	0.34	0.37	37.4	33.02
	Bagging MLP	**67.57**	**71.54**	**67.57**	**69.50**	**0.35**	**0.38**	**38.33**	**32.44**
	AdaBoost MLP	66.3	66.64	66.28	66.46	0.33	0.4	39.57	33.71
	Majority Voting DL+NN+MLP	66.69	67.52	66.69	67.10	0.43	0.37	36.51	33.31

Table 3: Average performance measure of Datasets 1 & 2

		Accuracy	Precision	Recall	F-Measure	Kappa	Absolute Error	Relative Error	Classification Error
Individual Classifiers	**DL**	74.25	72.28	69.37	70.80	0.39	0.3	30.32	26.25
	NN	80.54	80.03	75.65	77.78	0.54	0.22	21.6	19.46
	AutoMLP	80.58	80.02	76.08	78.00	0.54	0.23	22.73	19.17
Ensemble Classifiers	**Bagging DL**	70.76	57.84	61.47	59.60	0.29	0.31	30.5	29.25
	AdaBoost DL	73.9	70.46	73.74	72.06	0.09	0.3	29.84	26.1
	Bagging NN	79.92	81.56	77.3	79.37	0.46	0.22	21.89	18.33
	AdaBoost NN	80.02	77.49	77.34	77.41	0.52	0.23	22.56	19.97
	Bagging MLP	**80.86**	**81.88**	**75.28**	**78.44**	**0.54**	**0.23**	**23.19**	**19.14**
	AdaBoost MLP	80.08	78.03	75.56	76.78	0.53	0.24	23.85	19.92
	Majority Voting DL+NN+MLP	80.27	74.91	75.23	75.07	0.58	`0.22	21.87	19.73

Results acquired with dataset 2 are DL 57.08%, NN 67.15%, AutoMLP 67.24%, Bagging DL 50%, AdaBoost DL 56.72%, Bagging NN 67.9%, AdaBoost NN 66.99%, Bagging MLP 67.57%, AdaBoost MLP 66.3%, voting DL+ NN + AutoMLP 66.69%. Average results of dataset 1 and dataset 2 DL 74.25% ,NN 80.54% , AutoMLP 80.58%, Bagging DL 70.76%,AdaBoost DL 73.9%, BaggingNN 79.92% ,AdaBoost NN 80.02%, Bagging MLP 80.86%, AdaBoost MLP 80.08%, voting DL+NN+AutoMLP 80.27%. Among all we have concluded that although NN, AutoMLP, AdaBoost NN, Bagging MLP, AdaBoost NN, voting DL+ NN + AutoMLP have obtained 80.02 to 80.86 % accuracy and Bagging NN have got the 79.92 % accuracy but by analyzing the other performance measure factors we conclude that Bagging NN as our best algorithm which can work efficiently on large datasets for more accurate results with lesser error rate results as shown in table 2, 3 & 4.

The average performance measure graphs for accuracy, precision and classification error are presented in figures 3,4 and 5 respectively.

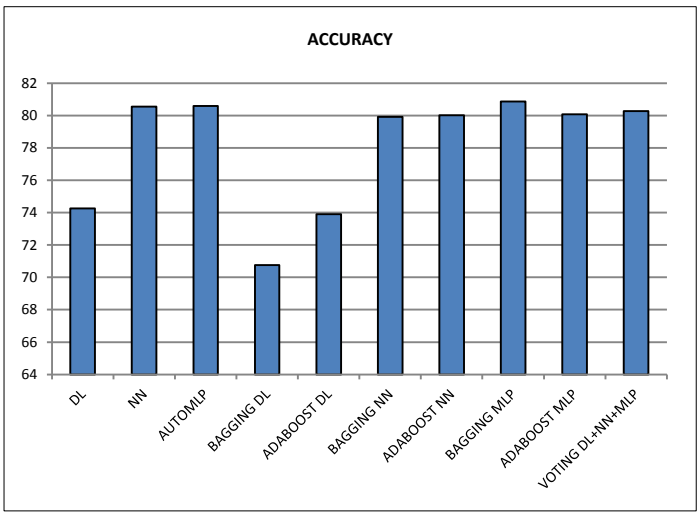

Figure 3: Accuracy Comparison

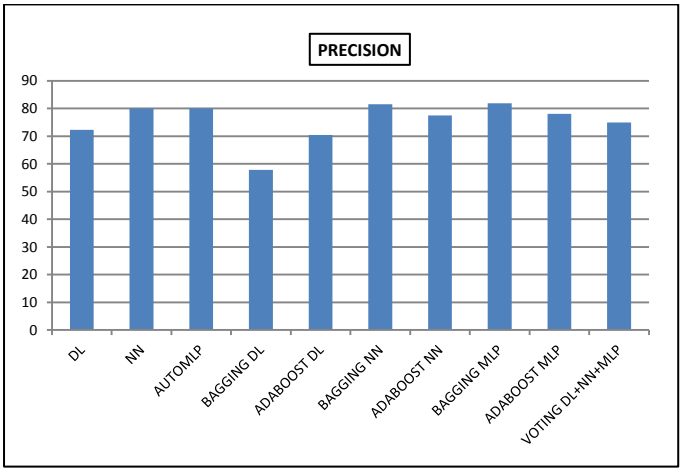

Figure 4: Precision Comparison

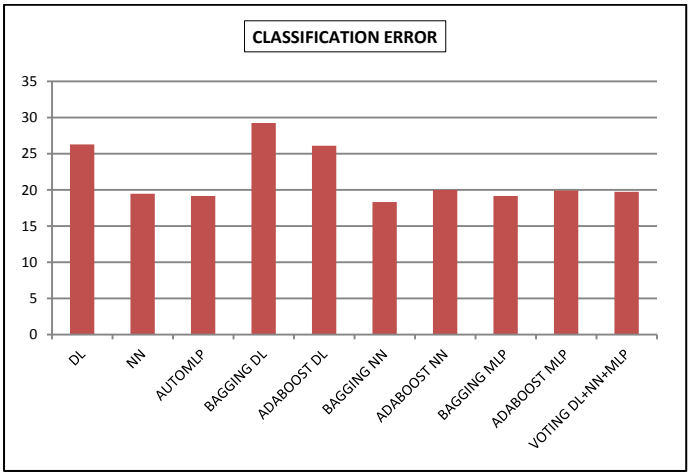

Figure 5: Classification Error Comparison

5. CONCLUSIONS AND FUTURE WORK

This research work proposed the Neural Network based Ensemble classifiers which are performing best for the given two datasets of telecom industry. These two datasets uses different types of attributes for the churn prediction so we are sure that the same model can be applied to any dataset for acquiring the best prediction result and can save the loyal customer of a company before churn. For future work we are proposing a new ensemble based on deep learning which may be used to recommend which combination of classifiers work best on the given type of datasets.

References

[1] Bi, W., Cai, M., Liu, M., & Li, G. (2016). A big data clustering algorithm for mitigating the risk of customer churn. IEEE Transactions on Industrial Informatics, 12(3), 1270-1281.

[2] Xia, G. E., Wang, H., & Jiang, Y. (2016, November). Application of customer churn prediction based on weighted selective ensembles. In Systems and Informatics (ICSAI), 2016 3rd International Conference on (pp. 513-519). IEEE.

[3] Dolatabadi, S. H., & Keynia, F. (2017, July). Designing of customer and employee churn prediction model based on data mining method and neural predictor. In Computer and Communication Systems (ICCCS), 2017 2nd International Conference on (pp. 74-77). IEEE.

[4] Franciska, I., & Swaminathan, B. (2017, May). Churn prediction analysis using various clustering algorithms in KNIME analytics platform. In Sensing, Signal Processing and Security (ICSSS), 2017 Third International Conference on (pp. 166-170). IEEE.

[5] Wang, C., Li, R., Wang, P., & Chen, Z. (2017, July). Partition cost-sensitive CART based on customer value for Telecom customer churn prediction. In Control Conference (CCC), 2017 36th Chinese (pp. 5680-5684). IEEE.

[6] Zhang, Z., Wang, R., Zheng, W., Lan, S., Liang, D., & Jin, H. (2015, November). Profit maximization analysis based on data mining and the exponential retention model assumption with respect to customer churn problems. In Data Mining Workshop (ICDMW), 2015 IEEE International Conference on (pp. 1093-1097). IEEE.

[7] Rodan, A., & Faris, H. (2015, November). Echo state network with SVM-readout for customer churn prediction. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

[8] Do, D., Huynh, P., Vo, P., & Vu, T. (2017, December). Customer churn prediction in an internet service provider. In Big Data (Big Data), 2017 IEEE International Conference on (pp. 3928-3933). IEEE.

[9] E. Shaaban, Y. Helmy, A. Khedr, and M. Nasr, "A Proposed Churn Prediction Model." Vol. 2, Issue 4, June-July 2012, pp.693-697.

[10] A. Idris, A. Iftikhar, Z. Rehman,"Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling." Cluster Comput DOI 10.1007/s10586-017-1154-3.

[11] A. Anjum, A. Zeb , I. Afridi, P. Shah et al ," Optimizing Coverage of Churn Prediction in Telecommunication Industry", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 5, 2017.

[12] Sahar F. Sabbeh,"Machine-Learning Techniques for Customer Retention: A Comparative Study", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018

[13] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354-2364.

[14] Mitkees, I. M., Badr, S. M., & ElSeddawy, A. I. B. (2017, December). Customer churn prediction model using data mining techniques. In Computer Engineering Conference (ICENCO), 2017 13th International (pp. 262-268). IEEE.

[15] Semrl, J., & Matei, A. (2017, October). Churn prediction model for effective gym customer retention. In Behavioral, Economic, Socio-cultural Computing (BESC), 2017 International Conference on (pp. 1-3). IEEE.

[16] Wenjie Bi, Meili Cai, Mengqi Liu, and Guo Li,"A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn," IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO. 3, JUNE 2016

[17] Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An Efficient Rule-Based Classification of Diabetes Using ID3, C4. 5, & CART Ensembles. In 2014 12th International Conference on Frontiers of Information Technology (FIT) (pp. 226-231). IEEE.