International conference on Advanced Techniques in Communication Networking and Automation

RESEARCH ARTICLE OPEN ACCESS

# Enhancing Customer Churn Prediction: A Hybrid Ensemble Approach with Clustering andClassification Techniques

Vivekanandhan. A
Assistant professor, CSE
A.V.C College of Engineering
Mannampandal, Mayiladuthurai
vivekcse@avccengg.net

Manikandan. S
SakthiMakesh. A
Sivasankar. M
Vignesh. A
B.E Computer Science and Engineering
A.V.C College of Engineering
Mannampandal, Mayiladuthurai
raksan0012@gmail.com

*Abstract*- **The customer churn prediction (CCP) is one of the challenging problems in the telecom industry. With the advent of machine learning models, the possibilities to predict customer churn has increased significantly. Although a number of algorithms have been proposed, there is still room for performance improvement. This study proposes a customer-churn prediction system that uses an custom built ensemble framework that is a hybrid model based on a combination of clustering and classification algorithms. First K-mediods clustering algorithm were evaluated. Then hybrid models were introduced by combining the clusters with four different classifiers called LSTM, SVM, Stochastic Gradient descent and logistic regression and then evaluations were performed using four ensembles techniques . The proposed research was evaluated on telecom datasets obtained from kaggle platform.**

**Keywords- Churn prediction, Classification, Clustering, Bagging, Boosting, Voting, Stacking.**

## I.Introduction

In the last two decades, the telecom industry has grown in significance, particularly in developed nations [1] This sector now provides a range of services aimed at retaining customers due to the rising number of telecom company users. The tendency of customers to stop doing business with a company in a specific time frame is known as customer churn. A precise and high-performance method for identifying clients who are likely to churn must be developed due to the importance of client retention. Customer retention is important as it allows companies to focus on keeping current customers happy, who may in turn refer new business. Additionally, keeping existing clients costs less than finding new ones. Long-term customers are loyal and less expensive to serve, and satisfied customers spread positive word-of-mouth [2].

Classification models are developed by training them on historical customer data, which is then applied to classify unseen patterns. In addition to single classifiers, ensembles like random forests and adaboost have shown to perform better than single classifiers in a number of studies. Ensemble techniques utilize multiple classifiers to acquire a better success than a single learner. An ensemble classifier is a set of consolidated weak classifiers. Ensemble methodology trains multiple classifiers on a given dataset, where each classifier is trained independently. In the first step, a diverse group of classifiers is generated from the training dataset, and in the second step, the outputs of the classifiers are consolidated to acquire a final decision. The main goal of ensemble methodology is to improve the performance of a single classifier by combining the outputs of multiple classifiers [3]. If all the classifiers are constructed using the same technique, the ensemble system is called homogeneous; If not, it is described as heterogeneous.

The proposed churn prediction model combines clustering and classification algorithms and is evaluated on a churn prediction dataset using metrics such as accuracy, precision, recall, and f-measure. The objectives of this research are to identify issues in existing literature and provide an efficient customer churn prediction model that accurately identifies potential churners for targeted retention strategies. The experiments demonstrate that the proposed model achieves high accuracy and outperforms other churn prediction models.

## II. Related Work

Reference [1], the authors presented the JIT-CCP model for predicting churn. The model consists of two main steps: data pre-processing and binary classification. The performance of the model is evaluated using the confusion matrix values of true positive, true negative, false positive, and false negative. The probability of detection (PD) is calculated based on these values, and it is used to assess the accuracy of multiple classifiers. If the PD value is close to 1, it indicates that the classifier's results are more accurate, and vice versa. However, the authors noted that the proposed model may not be suitable for handling a large volume of data.

Reference [4] highlights the importance of customer retention for the success of a company in the telecom industry. Retaining existing customers is more cost-effective than acquiring new ones, and it helps to maintain the company's ranking and increase its profits. To achieve this, customer association management (CAM) is crucial. The study proposes a hybrid model that uses both supervised and unsupervised techniques for churn prediction. The model involves several stages, including data cleaning, obtaining tes ting and training sets from different clusters, applying prediction algorithms, and evaluating the efficiency of the proposed model based on accuracy, specificity, and sensitivity.

Reference [5] discusses the application of a gravitational search algorithm for feature selection in customer churn datasets. The selected features were then used in ensemble classifiers such as random forest and support vector machines to predict customer churn. However, the gravitational search algorithm has certain limitations, such as slow convergence and a tendency to fall into local optima.

Reference [6] discusses the application of the firefly algorithm for predicting customer churn. The author used the firefly algorithm for feature selection and classification in the Orange dataset.

Reference [7] presents a customer-churn prediction system that utilizes an ensemble-learning technique consisting of stacking models and soft voting. The study selects four machine-learning algorithms, including Xgboost, Logistic regression, Decision tree, and Naïve Bayes, to build a stacking model with two levels. The outputs of the second level are then used for soft voting. To expand the feature space and uncover latent information from the churn dataset, equidistant grouping of customer behavior features is used for feature construction.

Reference [8] Author applied the PSO method for the feature selection and Feed Forward Neural Network for the prediction of customer churns. The oversampling method is applied to handle the imbalance dataset where the feed-forward neural network has the overfitting problem in the training process.

Reference [9] evaluates the performance of individual and ensemble neural network-based classifiers for churn prediction and proposes an improved ensemble classifier that uses bagging with neural network. The goal is to increase churn prediction's precision. The study employs two benchmark datasets obtained from GitHub to compare and evaluate the proposed model.

## III. Proposed methodology

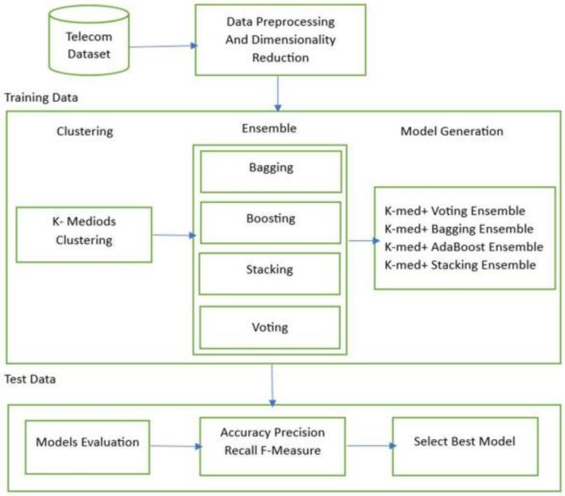

Figure 1 :Architecture of proposed model

Dataset:

Since the first objective involves gathering data, we chose a dataset from Kaggle, a free online data repository for research, for that challenge. This dataset's 7043 unique values and 21 distinct features provide data on the behavioural, demographic, and financial concerns of customers.

DATAPREPROCESSING

Data pre-processing's primary goal is to eliminate noise, abnormalities, missing values, and duplication from data. [10]. As this dataset is free of missing value, feature selection is implemented. The goal of feature selection is to reduce the number of features that are being used by the predictive model. Feature selection is performed to avoid the curse of dimensionality, which makes results more understandable, slows down the processing time, and lowers the predictive power of the models.

FEATURE SELECTION

Feature selection is performed using selectkBest and most important features are chosen for prediction model

CLUSTERING ALGORITHM

Following the data preprocessing, clustering has been utilized as a mean to enhance the predictive capabilities of the proposed model.

K-MEDOIDS CLUSTERING

A clustering method known as the K-Medoids algorithm, which is likewise partition-based, was introduced in 1987 by Rousseeuw Lloyd and Kaufman. In comparison to K-Means, K-Medoids is more resilient to noise and outliers [11]. The cost of each cluster is calculated using the formula below.

$$c = \sum_{c_i} \sum_{p_i \in c_i} |p_i - c_i|$$

where Pi and Ci are objects for which dissimilarity is calculated.

CLASSIFICATION ALGORITHMS

The proposed model carries out classification next to clustering. Out of which the most accurate churn prediction approach will be assessed through the combination of clustering and ensemble classification techniques.`

SUPPORT VECTOR MACHINE

SVM is a machine learning algorithm which tries to find the hyperplane that best separates the positive (churned customers) and negative (Non churned customers) samples in a high-dimensional feature space. The hyperplane is chosen such that the margin between the closest positive and negative samples (known as support vector) is maximized. The formula represents the hyperplane:

w ^ TX + b = 0

where w is the weight vector, b is the bias term, and X is the feature vector. The optimisation issue is shown as

min 1/2||w||^2+C*sum(max(0,1-Y_i*( w^TX_i + b)))

where ||w||^2 is the L2 norm of the weight vector, C is a regularization parameter, and max(0,1-Y_i*(w^TX_i + b)) is a hinge loss function that penalizes misclassifications.

LOGISTIC REGRESSION

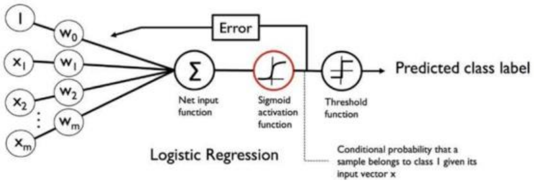

Figure 2 : LOGISTIC REGRESSION

Logistic Regression is a simple, well interpretable linear method which provides probabilistic predictions for identifying high risk customers and developing targeted retention strategies. It is based on the logistic function, which maps any input value to a value between 0 and 1. In churn prediction, the logistic function is used to model the probability of churn as a function of customer characteristics:

p(x) = 1/(1+ exp(-(b0 + b1x1 + b2x2+ ...+bnxn)))

where p(x) is the probability of churn for a customer with feature vector x, b0 is intercept, and b1,b2,...,bn are the coefficients that correspond to the n features in the model.

LONG TERM SHORT MEMORY

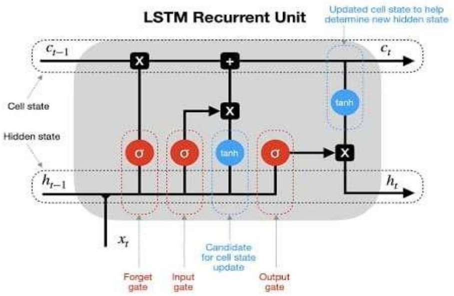

Figure 3 : LSTM

LSTM is a type of recurrent neural network (RNN) that is well suited for processing sequential data. In churn prediction, customer behavior over time is an important factor in predicting whether a customer is likely to churn or not.

The equations for the LSTM architecture are as follows:

ft = σ(Wf[h_t-1, x_t] + bf) (i)

it = σ(Wi[h_t-1, x_t] + bi)---- (ii)

Compute the forget gate that determines how much of the previous cell state to forget, based on the input `x_t` and previous hidden state `h_t-1`(Eq 1).Compute the input gate that determines how much of the candidate activation to add to the cell state, based on the input `x_t` and previous hidden state `h_t-1`(Eq 2).

C~t = tanh(Wc[h_t-1, x_t] + bc) ----(iii)

Compute the candidate activation that could be added to the cell state, based on the input `x_t` and previous hidden state `h_t-1`.(Eq 3)

C_t = ft * C_t-1 + it * C~t ---- (iv)

ot = σ(Wo[h_t-1, x_t] + bo) ---- (v)

Update the cell state by taking a linear combination of the previous cell state and the candidate activation, based on the forget and input gates.Compute the output gate that determines how much of the cell state to output, based on the input `x_t` and previous hidden state `h_t-1` (Eq 4,5).

h_t = ot * tanh(C_t) ----(vi)

Compute the current hidden state by taking the element-wise product of the output gate and the hyperbolic tangent of the cell state.(Eq 6)

STOCHASTIC GRADIENT DESCENT

SGD is an optimization algorithm used for minimizing the cost function. In SGD, the model parameters are updated using mini-batches of data rather that the entire dataset at once which results in faster convergence and reduced memory usage. The formula for updating the weights (parameters) in SGD is:

θ = θ - α * (∂J/∂θ)

where θ is the vector of parameters (weights), α is the learning rate, J is the cost function, and (∂J/∂θ) is the gradient of the cost function with respect to the weights.

ENSEMBLE CLASSIFIERS

[14] used ensemble methods to apply multiple learning algorithms for prediction. Ensembles increase the performance of the system or model [15].

VOTING

The voting or plurality method is employed when multiple classification algorithms are used, and it involves selecting the class label

with the highest frequency vote from each classifier.[12]

$$\sum_{t=1}^{T} d_{t,j}(x) = max_j = 1,2,3 \ldots c \sum_{t=1}^{T} d_{t,j}|$$

where T represents the number of classifiers, and d(t,J) is the decision of classifier and J represents the classes.

BAGGING

Bagging, short for Bootstrap Aggregation, is an ensemble classifier that uses a collection of similar and dissimilar objects. Its purpose is to reduce the variance of classifiers used in a prediction model, resulting in better performance [13].Then evaluation

of Bagging is given as follows:

$$V_{t,j} = \begin{cases} 1 & \textit{if } h_t \textit{ picks class } w_j \\ 0 & \textit{otherwise} \end{cases}$$

where t represents training samples, ht represents trained classifiers and wi represents class

labels. Each class will have total votes represented by:

$$v_j = \sum_{t=1}^{T} v_{t,j} = 1,2,3 \ldots . c$$

BOOSTING

XGBoost, or eXtreme Gradient Boosting, is a popular implementation of the gradient boosting algorithm that combines the strengths of multiple decision trees into a more accurate and robust prediction model. By iteratively training new trees to correct the errors of the previous ones, XGBoost can improve the prediction accuracy and reduce the bias and variance of the final model [7].

Obj = L + Ω

where L is the loss function, and $\Omega$ is the regularization term that penalizes the complexity of the model to avoid overfitting

$$L = - \sum[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

where the actual value of the is $y_i$

target variable for the i-th training example, and $p_i$ is the predicted probability of the positive class for the i-th example. The regularization term $\Omega$ is typically defined as the sum of the absolute values of the weights or the squared values of the weights.

STACKING

Stacking is an ensemble method that combines multiple learners rather than choosing the best one among them. This technique is useful for achieving better performance than any individual model. The training data is bootstrapped to train Tier-1 classifiers, which generate predictions that are used to train Tier-2 classifiers. This approach ensures that the training data is used effectively for learning.

| Technique | Accuracy | Recall | Precision | F-measure |
|---|---|---|---|---|
| LSTM | 86.02 | 74.5 | 84.6 | 71.42 |
| SVM | 88.6 | 75.28 | 81.88 | 72.06 |
| STOCHASTIC GRADIENT DESCENT | 91.51 | 84.8 | 82.1 | 78.44 |
| LOGISTIC REGRESSION | 89.7 | 83.12 | 78.46 | 80.63 |

References

[1] Ullah I, Raza B, Malik AK, Imran M, Islam SU, Kim SW. 2019. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factoridentification in telecom sector. IEEE Access 7:60134–60149

[2] Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. Eur. J. Oper. Res. 2012, 218, 211–229. [CrossRef]

[3] Gopika, D., Azhagusundari, B. (2014). An analysis on ensemble methods in classification tasks. International Journal of Advanced Research in Computer and Communication Engineering, 3(7):7423–7427

[4] Vijaya J, Sivasankar E. 2018. Improved churn prediction based on supervised and unsupervised hybrid data mining system. In: Information and Communication Technology for Sustainable Development. Singapore: Springer, 485–499

[5] Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., 2021. Customer churn prediction system: a machine learning approach. Computing,pp. 1-24.

[6] Ahmed, A.A. and Maheswari, D., 2017. Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal, 18(3), pp. 215-220

[7] Tianpei Xu, Ying Ma and Kangchul Kim 2021.Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping.Appl. Sci. 2021, 11, 4742.

[8] Faris, H., 2018. A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors. Information, 9(11), pp. 288.

[9] Mehpara Saghir et al.,2019 Churn Prediction using Neural Network based Individual and Ensemble Models.IBCAST

[10] Azeem M, Usman M, Fong ACM. 2017. A churn prediction model for prepaid customers in telecom using fuzzy classifiers. Telecommunication Systems 66(4):603–614 DOI 10.1007/s11235-017-0310-7.

[11] Gajowniczek K, Orłowski A, Ząbkowski T. 2019. Insolvency modeling with generalized entropycost function in neural networks. Physica A: Statistical Mechanics and its Applications 526(1):120730 DOI 10.1016/j.physa.2019.03.095.

[12] Gupta MK, Chandra P. 2020. A comprehensive survey of data mining. International Journal of Information Technology 12(4):1243–1257 DOI 10.1007/s41870-020-00427-7.

[13] Brown G, Wyatt J, Harris R, Yao X. 2005. Diversity creation methods: a survey and categorisation. Journal of Information Fusion 6(1):5–20 DOI 10.1016/j.inffus.2004.04.004

[14] Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M. 2017. Ensemble learning for data stream analysis: a survey. Information Fusion 37(2):132–156 DOI 10.1016/j.inffus.2017.02.004.

[15] Rustam F, Mehmood A, Ullah S, Ahmad M, Khan DM, Choi GS, On BW. 2020. Predicting pulsar stars using a random tree boosting voting classifier (RTB-VC). Astronomy and Computing 32:100404 DOI 10.1016/j.ascom.2020.1004