Volume 9, Issue 4, April – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24apr643

Clustering Comparison of Customer Attrition Dataset using Machine Learning Algorithms

Anitha R[1]; Aameer Khan S[2]; Harini Murugan[3]; Nithisshkrishna KS[4]

[1-4]Department of Artificial Intelligence and Machine Learning,
Rajalakshmi Engineering College, Chennai, India

Publication Date: 2025/03/19

Abstract: In the dynamic landscape of today's business environment, customer retention is a critical factor for sustainable growth and success. This project focuses on developing and comparing machine learning models for customer attrition and churn prediction using state-of-the-art algorithms such as Affinity, Birch, KMeans, and Agglomerative Clustering. The objective of this study is to evaluate the effectiveness of these clustering algorithms in identifying patterns and predicting customer churn. Using a dataset containing historical customer data, the project aims to create prediction models that can assist firms in proactively addressing possible churn concerns and implementing targeted retention efforts. The study is significant because it can give businesses predictive analytics capabilities to enhance their customer relationship management strategies, by figuring out which customers are likely to leave. In addition, the project intends to execute label selection by evaluating each feature individually according to its impurity score and to perform cluster classification to choose the optimal cluster according to its metrics. The study concentrates on the crucial machine learning methods for calculating client churn. This can include improving customer service, offering loyalty programs, or adjusting pricing strategies.

Keywords: *Customer Attrition - Apache Spark - K-Means Clustering - Web Application - Customer Retention- Logistic Regression - Machine Learning Algorithms.*

I. INTRODUCTION

Customers represent the utmost value in any industry resource since they are the primary engine of profit generation. Organizations in today's world understand that they should invest in a lot of strategies that encourage customer retention and satisfaction. For many years, businesses have utilized.

customer churn to increase revenue and create long-lasting relationships with their clients [16]. Churners are those who relocate to other companies for several reasons. To reduce customer turnover, the organization should possess the ability to accurately forecast the customer actions and underlying causes under their control. The binary classification task of prediction distinguishes churners from non-churners. Customer attrition prediction has become a critical area of focus which gained interest in recent times. Telecommunication companies, in particular, are facing increasing pressure to retain their customers in a saturated market. The emergence of customer attrition prediction has led to a significant change in the telecom industry. Churn is one of the most important service aspects in the telecommunications industry [9]. The customer data collected from multiple diverse contact points empowers companies to develop personalized products, nurture innovation, tailor products and services, and thereby enhance customer satisfaction and competitive advantage [2]. Due to their inability to forecast when a consumer will go, many businesses frequently experience customer loss. The main goal of this research is to provide telecom businesses with rapid and precise techniques for identifying consumers likely to leave. However, in the telecommunication industry, the customer churns quite often. The highly competitive nature of telecom industry companies actively monitors customer behavior and predict churn through data-driven insights and also allocate resources strategically to retain customers [3]. Customer churning is the analysis or estimate of analysis or the degree of customers who turn to shift to an alternative [12].

Reactive and proactive customer churn management are the two main approaches, as demonstrated by Van den Poel and Burez [1]. When a business adopts a reactive strategy, it holds off on terminating its service connection until the client requests it. In this instance, the business will provide the client a reason to stick around. When the business takes a proactive approach, it looks for clients who are about to purchase before the others do. The business then gives them unique incentives to prevent these customers from quitting. Employing machine learning techniques that learn from data iteratively: K-means clustering and logistic regression are two well-liked algorithms that can be used to forecast client attrition.An approach for supervised learning called logistic regression can be used to model the likelihood of a binary outcome, such as a customer's likelihood of churning. To calculate the likelihood of churn, a linear model is built using several independent

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24apr643

variables, including customer satisfaction, consumption trends, and demographics. An unsupervised learning approach called K-means clustering can be used to cluster data points according to how similar they are. To minimize the separation between data points within a cluster and increase the separation between data points in distinct clusters, iteratively divide the data into a predefined number of clusters. While there are numerous methods for predicting and assessing customer attrition, only a small number of them generate accurate forecasts and function well with large amounts of data.

II. METHODOLOGY

➢ *Random Forest Algorithm (Xiancheng Xiahou):*

The methodology utilized in this research is the Random Forest algorithm. This algorithm was employed as a robust and efficient feature selection method. Random Forest is widely recognized for its exceptional classification preciseness, capability to withstand noise and anomalies, and to generalize well across various domains including business management, economics, finance, and biological sciences.Given the dataset's considerable dimensionality of 17 variables, the challenge was to determine the number of features (M) to include in the predictive model. To address this, the Out-of-Bag (OOB) error was utilized as a standard of measurement for feature selection. In order to compute the OOB error, different bootstrap samples were used for the training set throughout the construction of every tree in the Random Forest. Surprisingly, the number of randomly selected features changed. It was seen that the distinctions in the OOB mistake rates were minimal.This suggested that the choice of the feature count (M) did not significantly impact the model's performance. Consequently, the decision was made to select four features in each iteration, resulting in a relatively low OOB error.This suggested that the choice of the feature count (M) did not significantly impact the model's performance. Consequently, the decision was made to select four features in each iteration, resulting in a relatively low OOB error. Fourvariables were identified as crucial for predicting customer churn: "Night Buy," "PM Buy," "Night PV," and "PM PV." These variables were considered key indicators for predicting customer loss in the churn prediction model [15].

➢ *U-Net (Karan Jakhar et al.):*

In the realm of data analysis, the abundance of available data often necessitates a process of classification, grouping this data into various categories or types, such as sound, video, and text designs. This characterization is fundamental for viable information mining, which envelops a scope of functionalities like grouping, segregation, affiliation, and bunching. Numerous complete frameworks are intended to give a set-up of information mining functionalities inside a solitary stage (Neha and Vikram, 2015). One notable classification technique is the Support Vector Machine (SVM), which excels in handling linear permutations of subsets within a training dataset. SVM aims to find a maximum margin separating hyperplanes in a high-dimensional feature space, which is particularly useful when dealing with nonlinearly separable information highlights (Nadeem, Umar, and Shahzad, 2018). This method effectively organizes data based on the most significant characteristics, even in scenarios where the vectors are nonlinearly separable [14]. In the SVM system, a few key parts assume essential parts:

- M: Indicates how many samples there are in the training set.
- Xi: Denotes vector support when the value of ai exceeds 0.
- 'X': Represents an unidentified vector sample.
- δ (delta): Serves as a threshold or margin.
- (ai): Parameter derived from solving a convex quadratic programming problem related to linear constraints.

In practice, various kernel functions are employed, such as the Polynomial kernel and Gaussian radial basis functions (RBF), to transform data into higher-dimensional spaces, permitting greater effectual class separation. The threshold (δ) is another parameter determined by selecting any 'i' where ai is greater than 0, and it satisfies the Karush–Kuhn–Tucker condition (Burges, 1998).

In summary, SVM is a powerful classification technique that maximizes the margin between data points in a high-dimensional space. It is an important technique in data mining and classification tasks since it is especially helpful when working with complex and nonlinearly separable data.

➢ *KNN (Prabadevi. B):*

This kind of boosting is called random gradient boosting. For every repeat, a subset of the training data is randomly (and without replacement) chosen from the whole preparation dataset. Then, the randomly chosen subsample is used to fit the base student instead of learning from the complete example. A few possible stochastic variants are as follows: Once the columns have been subsampled, create each tree. Before creating each tree, subsample the segments.

➢ *Rule of Training.*

- The initial phase of the teaching for training.
- Inputs
- A tendency
- The learning rate should be set to a level appropriate for basic estimation, and NN parameters like biases and weights should be set to a desirable zero.
- Begin each information unit with the following Si (i=1 to n) = xi
- After the result, get feedback from the web.
- Using the appropriate activation function, ascertain a conclusive result based on the results of step 6.

In this paper, we introduce stochastic gradient boosting to enable the gradient boosting approach to be used for both continuous target variables and categorical objective variables (as a classifier or regression). The model's bias error is reduced by the application of gradient boosting. Log

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24apr643

loss is the price performed when it is used as a classifier, and mean square error (MSE) is the price performed when it is used as a regressor. Because it can improve a wide range of unlucky works and provides a few hyper boundary tuning options that make the capacity fit entirely flexible, stochastic gradient boosting is incredibly versatile. There is no need to prepare the data; both the numerical values and the categories function well [7].

➢ *Churn Prediction Using Naive Bayes (Khulood Ebrah):*

A set of conditional independence presumptions and the Bayes rule serve as the foundation for the Naive Bayes algorithm, which is a classification algorithm [11]. For each class Ci, P(X|Ci) P(Ci) is evaluated in order to predict the class label of X.The classifier predicts that the class label of tuple X is the class Ci if and only if.

- P(X|Ci)P(Ci)>P(X|Cj)P(Cj) for $1\leq j\leq m, j\neq i$ Stated differently, the class Ci with the highest P(X|Ci)P(Ci) is the projected class label [12]. models posterior probability using the Bayesian approach. Specifically, for every k=1,⋯, K,

- $P\hat{}(Y=k|X1,\cdots,Xp)=\pi(Y=k)\pi P(Xj|Y=k)/\sum\pi(Y=k)P(Xj|Y=k)$where Y represents the random variable associated with the churn class index of an observation. The predictors of an observation are X1, ⋯, Xp. (Y=k) is the previous probability that a class index is k.The model uses the mean and standard deviation to distribute predictors within each class.

Using Naive Bayes classification, the method determines the parameters of a probability distribution, provided that predictors are conditionally [4] independent given the class. Step 1: Predicting The method computes the posterior probability of a sample belonging to each class for any unseen test data. The test data is then categorized by the method according to the largest posterior probability.

➢ *Customer Churn Analysis Using LSTM-RNN Model (Nagaraju Jajam):*

Churn describes the consumer who transfers from one supplier of telephone services to another [17]. In order to precisely ascertain the probability of customer churn from the provided dataset, the LSTM-RNN model is applied in the churn classification process. In order to do this, a deep learning framework with an attention layer that improves the comprehension of churn classification accuracy is used. To fully execute the suggested LSTM-RNN model, a few more processing steps are needed. First, features are convolution, and then input data is loaded into the LSTM-RNN architecture. This stage is dedicated to obtaining detailed semantic information from the word order. Furthermore, the temporal relationships between features are efficiently identified and captured by the LSTM-RNN architecture, which in turn produces a feature vector that aids in the overall churn classification process. (5)

Assuming the semantic meaning of input data also entails comprehending the context and underlying information contained in the data, especially as it relates to customer behavior. Creating labels that indicate whether a customer has churned (assigned a value of 1) or not (assigned a value of 0) is an essential step in training such a model. It is recognized, nevertheless, that the sechurn label assignments may be somewhat arbitrary, given that churn is frequently determined by a variety of variables and interpretations.

A matrix layer is subsequently created from these labeled data points, and this layer is used as the input for a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). A particular kind of RNN cell called an LSTM cell is made to handle sequential data and crecognize long-range dependencies. Even though they work incredibly well for things like time-series analysis and natural language processing, they do have certain disadvantages. Because of their intricate architecture with many parameters and operations, LSTM cells have a higher computational cost and require more memory and time to train and operate. LSTM-RNNs are preferred because of their superior capacity to represent sequential data in spite of these disadvantages.

Simple RNN cells, on the other hand, require less computing power but have difficulty identifying long-term relationships in data. RNNs have benefits of their own within the larger neural network context. They are appropriate for a variety of tasks because of their architecture, which is based on deep neural networks and enables them to process information both sequentially and concurrently. Furthermore, the network can replicate some aspects of the brain's processing capabilities, particularly in retaining and utilizing information over lengthy sequences, thanks to the addition of memory cells, as seen in LSTM-RNNs. As a result, even with their high processing requirements, LSTM cells are still a good option for issues like churn prediction where it's important to comprehend sequential data and long-term dependencies.

➢ *Automated Pneumothorax Detection and Quantification from CT Scans (Soumi De):*

The Sampling-based Stack Framework (SS-IL) that has been proposed provides a new method for churn prediction. This framework makes use of ensemble learning to improve classifier performance. The outputs of several base classifiers are combined using the potent technique of ensemble learning to arrive at a final classification. A particular type of ensemble learning called stacking uses multiple base learners, also called level-0 learners, who are trained with the same training dataset.

The SS-IL framework is unique in that it uses different training datasets for the classifiers at level 0 of the classification. By using sampling techniques, the goal is to increase the variety of attributes taken into account and make it easier for the ensemble to gather important information. The goal of this training data diversification is to raise the framework's overall predictive power.

Furthermore, a meta-learner—an additional component of the SS-IL framework—is trained with the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24apr643

predictions produced by the level-0 learners. Effective instance classification is made possible by this meta-learner's acquisition of the combination weights for each of the decision probabilities supplied by the base-level classifiers. A stacked ensemble such as SS-IL relies on the level-0 base learners to facilitate the information gained from the features used in training the meta-learner. This framework is essentially based on the idea that multiple classifiers' combined knowledge and the variety of training data improve predictive robustness and accuracy.

Note that although the SS-IL framework is discussed in the context of churn prediction in this content, there are indications of possible medical applications as well, particularly for monitoring and diagnosing pneumothorax in clinical settings. The framework highlights its significance beyond predictive analytics by demonstrating its versatility and utility across different domains, potentially improving patient care and saving time.

III. PROPOSED IMPLEMENTATION

The paper aims to deploy an ensemble approach to provide businesses with a holistic and proactive solution for predicting customer churn and optimizing customer retention strategies.The methodology involves preprocessing and enhancing the dataset to ensure optimal performance of the predictive models. The customer base is segmented using Agglomerative Clustering, Affinity, Birch, and KMeans algorithms, enabling the identification of distinct customer groups with varying churn probabilities. Subsequently, predictive models are developed for each cluster, enhancing the granularity of churn predictions.The ensemble approach integrates the predictions from individual models, capitalizing on the strengths of each algorithm. This ensures a more robust and accurate prediction by considering diverse customer behaviors and patterns.

IV. RESULTS & DISCUSSION

The goal of business studies in the telecommunications sector is to increase their financial gains. Churn prediction is widely recognized as the primary revenue stream for telecommunications companies. This paper examined methods for developing a big data application that predicts the percentage of customer attrition. The telecom sector can benefit greatly from the combination of big data technology and machine learning to predict customer attrition. We understood how crucial it was to preprocess and get the data ready for analysis. Our goal was to identify the key elements influencing customer churn by analyzing pertinent attributes and customer behavior.Additionally, by employing methods like logistics regression and K-means clustering, significant attributes can be extracted from massive amounts of telecom data, and pertinent data can be fed into machine learning algorithms to anticipate and prevent customer attrition. All things considered, the approaches under review hold promise for developing a well-rounded machine-learning model that will aid the sector in reducing customer attrition.The industry's bottom line and customer satisfaction levels could both benefit greatly from these strategies. Telecom companies can improve customer satisfaction, lower revenue loss associated with churn, and ultimately bolster their competitive position in the market by precisely identifying customers who are at risk of leaving and customizing retention efforts to meet their specific needs. Our research aims to enhance comprehension of customer attrition prediction and stimulate additional investigation and creativity in the big data analytics and telecommunications domains.

REFERENCES

[1]. Burez J., & Van den Poel, D "Crm at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services", Expert Systems with Applications 32, 277– 288.
[2]. Ledro, C., Nosella, A., & Vinelli, A. (2022). Artificial intelligence in customer relationship management: literature review and future research directions. Journal of Business & Industrial Marketing, 37(13), 48-63.
[3]. Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, 101-112.
[4]. Khulood Ebrah, Selma Elnasir "Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms".Journal of Computer and Communications > Vol.7 No.11, November 2019.
[5]. Nagaraju Jajam, Nagendra Panini Challa, Kamepalli S.L.Prasanna "Arithmetic Optimization With Ensemble Deep Learning SBLSTM-RNN-IGSA Model for Customer Churn Prediction" in IEEE vol 11.
[6]. Soumi De, Prabu.P" A Sampling-Based Stack Framework for Imbalanced Learning in Churn Prediction in IEEE vol 10.
[7]. Prabadevi.B, Shalini.R, Kavitha.B.R (2023). Customer Churning analysis using machine learning algorithms. In International Journal of Intelligent Networks.
[8]. M. Alizadeh, D. S. Zadeh, B. Moshiri and A. Montazeri, "Development of a Customer Churn Model for Banking Industry Based on Hard and Soft Data Fusion," in IEEE Access, vol. 11, pp. 29759-29768, 2023, doi: 10.1109/ACCESS.2023.3257352
[9]. Anand, M., Shaukat, I., Kaler, H., Narula, J., & Rana, P. S. Hybrid Model for the Customer Churn Prediction
[10]. Zadoo, A., Jagtap, T., Khule, N., Kedari, A., & Khedkar, S. (2022, May). A review on churn prediction and customer segmentation using machine learning. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (Vol. 1, pp. 174-178). IEEE..
[11]. Mitchell, T.M. (2015) Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression.
[12]. Han, J., Pei, J. and Kamber, M. (2011) Data

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/24apr643

Mining: Concepts and Techniques. Elsevier, Amsterdam.
[13]. PM, U., & Balaji, N. V. (2019). Analyzing Employee attrition using machine learning. Karpagam Journal of Computer Science, 13, 277-282.
[14]. Abdulsalam Sulaiman Olaniyi , Arowolo Micheal Olaolu , Bilkisu Jimada- Ojuolape , Saheed Yakub Kayode,," Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithm. In International Journal of Multidisciplinary Sciences and Advanced Technology Vol 1 No 1 (2020) 48–54.
[15]. Xiancheng Xiahou and Yoshio Harada , "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM.
[16]. Seymen, O. F., Dogan, O., & Hiziroglu, A. (2020, December). Customer churn prediction using deep learning. In International Conference on Soft Computing and Pattern Recognition (pp. 520-529). Cham: Springer International Publishing.
[17]. Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in the telecommunication industry using deep learning. Information Sciences Letters, 11(1), 24.