Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Customer Segmentation in IT Sector using Datamining Techniques

T.Kalai Selvi[1], S.Sasirekha[2], N.Deepika[3], V.Kanagalakshmi[4], R. Kavya[5]

[1]Associate professor, Department of CSE, Erode Sengunthar Engineering College, Perundurai, Erode.

[2]Associate professor, Department of CSE, National Institute of Technical Teachers Training and Research, Chennai.

[3,4,5]Student, Department of Computer Science and Engineering, Erode Sengunthar Engineering College, Perundurai, Erode, Tamilnadu, India.

Email: [1]tkalaiselvi1281@gmail.com, [2]sasirekha@nitttrc.edu.in, [3]ndeepikagobi@gmail.com, [4]kanagasaru09@gmail.com, [5]kavyaramesh0903@gmail.com.

Abstract

Due to its large client base, the IT industry generates enormous amounts of data every day. Business experts and decision-makers stressed that keeping current clients is less expensive than acquiring new ones. Business analysts and customer relationship management (CRM) analysts must understand the causes of customer attrition as well as the patterns of behavior found in the data of these clients. This research is a comprehensive study about churn prediction in IT industry it also suggests a churn prediction model to identify consumer churn and provide the reasons behind customer churn in the IT industry by employing clustering and classification algorithms. Information gain and correlation attribute ranking filters are used in feature selection. One of the CRM's most important tasks is to create retention strategies that work to keep customers from leaving. The study intends to create a model that can effectively identify the primary reasons behind customer churn. This model is likely to use various techniques, such as analyzing customer behavior, preferences, or other relevant data, to identify patterns or characteristics associated with churn and segment the churning customers based on groups in order to retain the customers based on the specific characteristics or behaviors they

DOI: https://doi.org/10.36548/jaicn. 2024.1.002
Received:14.12.2023, received in revised form: 16.01.2024, accepted: 02.02.2024, published: 13.02.2024

share within their respective groups. The main aspect of the study is to identify the challenges in the existing methods that are used in churn prediction and suggest a model to improve churn prediction and elude churn effectively.

Keywords: Association Rule Mining, Customer Segmentation, Market Analysis, IT Sector

1. Introduction

In today's rapidly evolving world, IT companies are confronted with an immense influx of data generated at an unprecedented pace. With a multitude of IT service providers vying for market share, customers are presented with numerous alternatives offering improved services at lower costs. The primary objective for IT companies is to maximize profitability and maintain competitiveness in the market. Predicting churn is critical in this industry, as operators aim to keep key clients and improve their Customer Relationship Management (CRM) tactics. Retaining existing clients is a huge problem, especially in a saturated and intensely competitive market when customers have the option to transfer suppliers. As a result, IT organizations have adopted techniques to discover and keep consumers, realizing that it is more cost-effective than acquiring new ones [10-12].

Classification groups data into predefined classes based on input features using supervised learning (e.g., Decision Trees, SVM). Clustering groups similar data points together into clusters without predefined class labels using unsupervised learning (e.g., K-means, hierarchical clustering). Information gain measures the reduction in uncertainty (entropy) of the target variable when splitting data based on a feature, and this helps in selecting the most informative features for decision-making and it used in decision tree algorithms. Correlation measures the strength and direction of linear relationship between two variables. It Ranges from -1 to 1, indicating negative, positive, or no correlation, and this helps in understanding associations between variables, but does not imply causation [13-15].

The proposed study aims to present a comprehensive review of the impact of mining and clustering techniques on predicting customer churn in different business sectors and suggests a model to predict customer churn in IT sector effectively and elude customer churn.

2. Related Study

The author in this study targeted marketing strategies focusing on market segmentation in retail, highlighting the limitations of conventional models. It employs Hadoop's distributed file system for data processing and conducts segmentation experiments using modified regression and clustering algorithms like EM and K-Means++.The research outcomes are two-fold: unveiling insights into Customer Lifetime Value (CLTV) segments and assessing the clustering algorithms' performance. Findings reveal a short customer lifetime (two years) with a high churn rate of 52%, leading to a devised marketing strategy that boosted departmental store sales growth from 5% to 9%. Customer relationships are deemed crucial and impacted by recent economic and social shifts, prompting a need for strategic planning and understanding consumer behavior and competitors. The study introduces new CLTV techniques and three RFM variations (P, Q, and T) offering advantages over standard models. These variations aid in identifying customer segments based on spending behavior, product attractiveness, and sales potential. Clustering algorithms like K-Means++ and EM, alongside the modified regression algorithm and RFMT variable, play a pivotal role in creating computable dynamic parameters, enhancing customer trends' understanding. Moreover, the study stresses the importance of accurate results obtained through the CLTV matrix against the RFMPQ dataset, emphasizing the application of marketing strategies to enhance customer satisfaction and loyalty. Age and gender variables contribute significantly, yielding an estimated analysis accuracy of 75% and resulting in a 6% sales growth rate [1].

The customer behavior analysis and segmentation using Data Mining (DM) methods, particularly in conjunction with the Regency Frequency Monetary (RFM) model is focused in the article. Its focus is to propose a customer segmentation framework by synthesizing various DM methods, emphasizing the significance of understanding customer characteristics for targeted marketing strategies. The study spans a review of articles from 2015 to 2020, culminating in the selection of 44 articles through specific criteria, predominantly focusing on RFM-based models and DM methods for segmentation based on empirical data. Association Rule Mining (ARM) and Apriori algorithms are pivotal in revealing frequently occurring patterns in customer segments and identifying item relationships. Furthermore, it proposes a framework incorporating DM and RFM-based CS within a Geographic Information Systems (GIS) environment for enhanced analysis and understanding of customer behavior [2] .

The author in this study emphasizes the shift from focusing solely on attracting new customers to retaining existing ones due to lower costs. It proposes an integrated customer analytics framework encompassing six components: “Data preprocessing, exploratory data analysis, churn prediction, factor analysis, consumer segmentation, and behavior analytics”. Three datasets and various machine learning classifiers are employed, with emphasis on imbalanced dataset handling using SMOTE and model evaluation through accuracy and F1-score. The study also stresses the significance of using SMOTE for imbalanced datasets and discusses better evaluation metrics. Furthermore, it calculates the overall churning probabilities for each cluster across the datasets, assisting in the prioritizing of retention efforts. This comprehensive approach, combining prediction of churn as well as segmentation for the telco industry, fills a gap in existing literature, contributing to a more holistic understanding of customer retention strategies [3].,

The author has proposed studya customer segmentation through data mining techniques, underlining the importance of understanding customer needs and preferences in marketing. It highlights the significance of computational analysis in uncovering hidden knowledge about customer behavior, aiding companies in delivering personalized services and fostering customer loyalty. Customer segmentation is emphasized as vital for market analysis, requiring proper data classification. The challenge lies in selecting suitable data mining techniques tailored to business needs and trends. The review discusses the challenge of selecting segmentation techniques, stressing the importance of considering management objectives and market trends. It presents four types of segmentation techniques: unobservable variables, general observable variables, product-specific unobservable variables, and product-specific observable variables Overall, the studystudy underscores the crucial role of data mining in customer segmentation, offering insights into methods for understanding and predicting customer behavior to aid businesses in targeting customers effectively and positioning their brands [4].

The study aims to manage massive volumes of GNSS (Global Navigation Satellite System) user connections efficiently by leveraging big data architecture and data mining algorithms. Python, integrated with Apache Spark, MongoDB, Matplotlib, and Folium libraries, serves as the core framework for data processing, querying, analysis, and visualization. The research presents a comprehensive approach that doesn't rely on proprietary

software, utilizing open-source tools for continuous real-time analysis. MongoDB, chosen for its open-source nature and versatile integration capabilities, ensures replication, high availability, and indexing based on navigation parameters. Hadoop, in combination with MapReduce, facilitates swift data recovery and analysis over vast user datasets, outperforming traditional proprietary GNSS software. The study highlights the superiority of data mining implementation in terms of speed, hardware requirements, and output information compared to other software solutions. Techniques like linear regression and Spark MLLib aid in analyzing user trends, economic recoveries, and seasonal correlations, especially within the agricultural sector. The study also sheds light on customer churning analysis and provides insights into connection frequencies, geographical demand, and clustering in real-time, offering comprehensive evaluations using a wide range of input and output formats [5].

The author in this research has proposed churn prediction, crucial in industries like telecom, banking, and life insurance. It explores ensemble-based classifiers (Bagging, Boosting, Random Forest) and compares them with established classifiers (Decision Tree, Naïve Bayes, SVM) for churn prediction in the telecom sector. The study emphasizes the significance of early detection of potential customer departure, particularly in telecom, employing machine learning and AI advancements for more accurate predictions. Experimental results reveal that Random Forest outperforms other models, boasting a 91.66% accuracy, low error rate, and high sensitivity. The telecom industry faces challenges in predicting customer departures, making churn prediction a fundamental concern, and this study delves into effective prediction models for this issue. The design of churn prediction models necessitates historical customer behavior data, typically gathered from various attributes, aiding in predicting future behavior. The dataset used, retrieved from a specific link, consists of 3333 records with attributes such as service calls, charges, and call durations, with churn as the class label. Additionally, the study outlines the confusion matrix parameters (True Positive, True Negative, False Positive, False Negative) for both ensemble-based and basic classifiers, clarifying their interpretations [6].

The study highlights the challenge the companies face in retaining customers, emphasizing that retaining existing customers is cost effective than acquiring new customers. The study proposes an ensemble-based framework utilizing various machine learning algorithms such as “k-NN, logistic regression, naïve Bayes, support vector machine, decision

tree, random forest, and multilayer perceptron". Experimental results reveal that the random forest algorithm achieves improved accuracy in predicting churns in all the sectors with the application of feature extraction techniques. It underlines the competitive market environment across telecom, banking, and insurance, stressing the importance of customer satisfaction in minimizing churn and retaining valuable customers. The study concentrates on homogeneous classifier ensembles, showcasing the superiority of the random forest method over traditional machine learning techniques across all sectors. Additionally, it discusses the future direction of applying deep learning methods to enhance churn prediction using extended datasets and diverse feature extraction techniques [7] .

The author in this research claims that the exponential growth of technology and the increased number of rivals in the telecom sector, organizations are suffering a severe problem with customer turnover. The problem of customer churn prediction (CCP) in the telecom industry, particularly for organizations with insufficient historical data. It presents a Just-in-Time (JIT) solution for CCP, based on cross-company data from another telecom operator. Empirical examination of this strategy using publicly available datasets from two telecom firms demonstrates that adopting a heterogeneous ensemble-based JIT-CCP model outperforms individual classifiers or homogeneous ensemble techniques. The study examines the performance of Support Vector Machine (SVM) as a base classifier in the proposed JIT-CCP model using homogeneous and heterogeneous ensemble approaches. The future study direction involves expanding the application of the proposed JIT-CCP model beyond telecommunications companies. The term "homogeneous" in this context denotes an approach utilizing a base classifier iteratively for diversity, exemplified by a bagging approach with SVM as the base classifier. The study specifies iteration and sample ratio values (e.g., 5 iterations and a 0.9 sample ratio) for training the SVM-based homogenous ensemble model.

The author in this study states that the telecoms business is extremely concerned about customer churn owing to dissatisfaction with service. Customer churn prediction in the telecommunications business is proposed using a unique feature selection (FS) approach called ACO-RSA, which combines ant colony optimization (ACO) and the reptile search algorithm (RSA). This ACO-RSA technique seeks to identify critical feature subsets for churn prediction, hence improving model performance and computing efficiency. The study of seven public churn prediction datasets and 10 CEC 2019 test functions demonstrates ACO-RSA's

superiority over other metaheuristic algorithms like PSO, MVO, GWO, standard ACO, and standard RSA. The telecommunications sector commonly employs churn prediction models using machine learning, necessitating effective feature discrimination. ACO-RSA, combining standard ACO and RSA, prioritizes exploration and exploitation, mitigating local optima. Despite its effectiveness, ACO-RSA requires slightly higher computational time (CT) during training to determine the best feature combination. The authors plan to optimize this by implementing standard ACO and RSA concurrently to reduce CT. The authors aim to apply ACO-RSA in diverse domains like renewable energy, IoT, and signal processing in future research [10].

Without adequate research and forecasting, the author has provided a mechanism for enterprises to identify repeatedly churning clients, in the telecom industry. The crucial issue of customer churn in the telecom industry is identified by employing the Deep-BP-ANN model, integrating two feature selection techniques, Variance Thresholding and Lasso Regression. Providers face immense pressure in delivering high-quality services like Audio, Video, and Internet access, necessitating a thorough understanding of customer churn dynamics. The study delves into the telecom sector's complexities, noting the numerous factors that may lead to customer churn, ranging from unsatisfactory service experiences to competitor offerings with better quality, price, or technological advancements. Analyzing such data becomes pivotal in predicting and preventing churn, enabling companies to address underlying issues that might cause customer dissatisfaction. It employs Lasso regularization for feature selection, notably impactful for the dataset Cell2Cell due to its larger feature set. These findings emphasize the significance of identifying and understanding key attributes influencing customer churn to aid in proactive churn prevention strategies [9].

Table 1. Comparative Table

Ref No	Dataset Used	Methodology	Demerits
[1]	RFMPQ	Bestfit Regression, K-means ++	Lacking sufficient market insights, Failing to identify sufficient small segments

[2]	Scopus,We of Science(WoS),E merald.	RFM-based model	Generalizability and Potential bias
[3]	Telecom, Cell2Cell, Cross Company	Bayesian logistic regression,K-means	Challenges in pratical implementation Framework's effectiveness may vary
[4]	Marketing dataset	Supervised and unsupervised data mining techniques	Over-reliance on data, Model Accuracy and Generalization
[5]	Geographic dataset	Deep-Bp-ANN Model	Potential Biases
[6]	Cell2Cell, Telecom	Ensemble-based classifiers(Bagging,Bo osting,Random forest)	Doesn't implementing essemble classifiers
[7]	Larger-Scale call Center,Insurance, Banking	Machine Learning Techniques	Using ensemble techniques potential limitations
[8]	Cross Company	SVM Method	Limited investigation into privacy implication, potential data sharing concerns
[9]	IBM Telco and Cell2cell	Deep-BP-ANN model	Potential biases
[10]	Public, Business Sector	ACO-RSA Method	High Computational Time

3. Proposed System

In the IT industry, a data-driven method that identifies churning clients and the variables contributing to their churn is the recommended churn prediction model. To achieve its objectives, the model combines clustering and classification methods. To ensure it utilizes the most relevant features for forecasting churn, the model initially conducts feature selection using information gain and correlation attribute ranking filters. Subsequently, the model

employs multiple effective classification algorithms, including Random Forest, Naive Bayes, Multilayer Perceptron, XGBoost, LightGBM, Logistic Regression, SVM, RNN, LSTM, etc., to classify churn customer data. After classification, the model utilizes similarity measures, the selection of which will be decided in the future based on the collected dataset. This is done to partition the churning customer data into groups

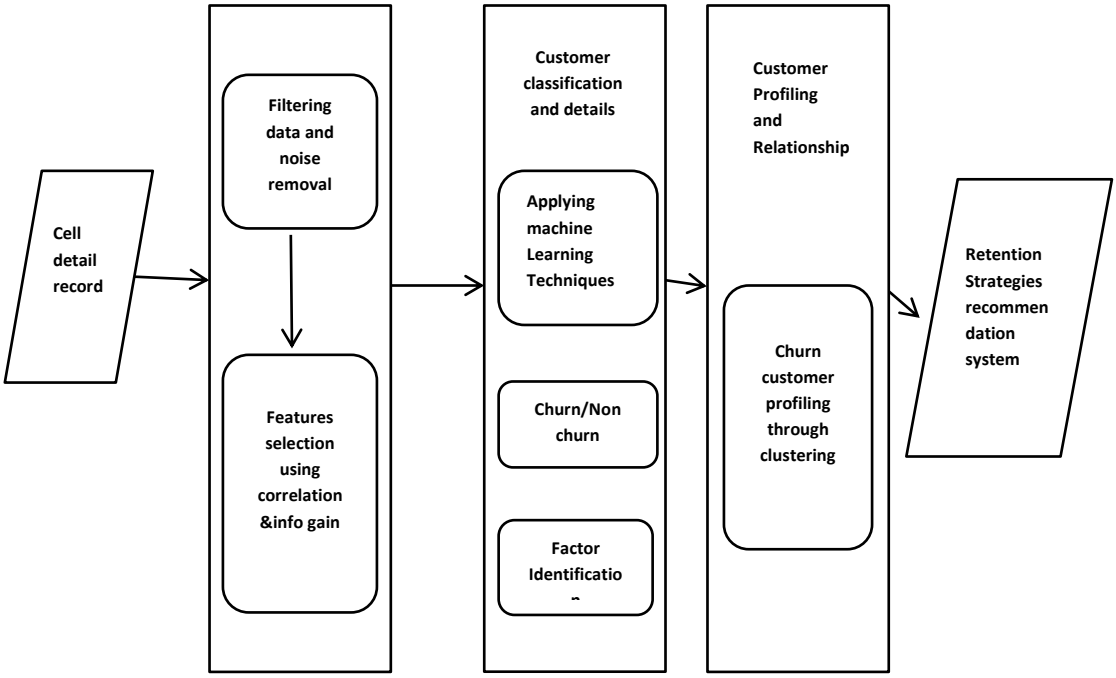

Figure 1. Block Diagram

4. Future Work

As data collection plays a crucial role in achieving accurate churn prediction, the proposed method for future research involves gathering data from diverse sectors of the IT industry in different states of Tamil Nadu. This will be done using various tools such as Google Forms, surveys, customer support interactions, customer feedback platforms, and sales and financial data. The data will be collected based on customer demographics, usage patterns, satisfaction levels, and other relevant factors. Subsequently, the collected data will undergo preprocessing to identify and handle unwanted and repeated information. The processed data will then be utilized in the proposed method to enhance churn prediction and effectively mitigate churn in the IT sector in Tamil Nadu. The future work will also encompass the

implementation, training, and performance evaluation of the suggested method to assess its efficiency and Accuracy.

5. Discussion

IT companies face the challenge of retaining customers amidst a highly competitive market. To address this, a data-driven churn prediction model is crucial. This model combines clustering and classification methods to identify churning clients and factors contributing to churn. By employing feature selection techniques such as information gain and correlation attribute ranking, the model ensures the use of relevant features for accurate churn forecasting. Classification algorithms applied to classify churn customer data. Following classification, similarity measures are utilized to group churning customers. This approach allows companies to target at-risk customers with special offers or promotions, reducing churn rates. Additionally, by proactively addressing customer service issues highlighted by the model, IT companies can improve overall customer satisfaction and loyalty. Moreover, insights from the churn prediction model enable companies to develop tailored products and services that meet the specific needs of their most loyal customers. In essence, the churn prediction model empowers IT companies to effectively manage churn, enhance customer satisfaction, and drive business growth through data-driven decision-making.

6. Conclusion

The implementation of a data-driven churn prediction model in the IT industry is instrumental in addressing the challenge of customer retention. By leveraging advanced clustering, classification, and feature selection techniques, IT companies can accurately identify churning clients and the underlying factors driving churn. This enables proactive measures such as targeted offers and improved customer service to mitigate churn and enhance overall customer satisfaction. Furthermore, insights from the churn prediction model facilitate the development of tailored products and services, fostering stronger customer relationships and loyalty. Ultimately, the adoption of such data-driven approaches empowers IT companies to stay competitive in a dynamic market landscape, driving sustained growth and success. The proposed study offers a comprehensive survey of the impacts of data mining on churn segmentation in the IT industry. It identifies potential limitations in the real-time application

of these methods and proposes a model to enhance churn prediction. The proposed model employs multiple classification algorithms and similarity measures to group churning customer data effectively. Implementation and performance measurement of the suggested model will be conducted in the future using an appropriate dataset. This dataset is planned to be collected from various IT sectors in different states of Tamil Nadu, aiming for an improved churn prediction and proactive churn prevention in the IT sector.

References

[1] Yoseph, Fahed, Nurul Hashimah Ahamed Hassain Malim, Markku Heikkilä, Adrian Brezulianu, Oana Geman, and Nur Aqilah Paskhal Rostam. "The impact of big data market segmentation using data mining and clustering techniques." Journal of Intelligent & Fuzzy Systems 38, no. 5 (2020): 6159-6173.

[2] M. Chen and Y. Hao, “A review of data mining methods in RFM-based customer segmentation,” IEEE J. Sel. Areas Commun., vol. 36, no. 3, pp. 587–597, Mar. 2021.

[3] R. Roman, J. Lopez, and M. Mambo, “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business,” IEEE Commun. Mag., vol. 78, no. 2, pp. 680–698, Jan. 2021

[4] H. El-Sayed et al., “Customer Segmentation via Data Mining Techniques: State-of-the-Art Review,” IEEE Access, vol. 6, pp. 1706–1717, 2021.

[5] Kumari, S. Tanwar, S. Tyagi, N. Kumar, R. M. Parizi, and K. R. Choo, “Big data architecture and data mining analysis for market segment applications of differential global navigation satellite system (GNSS) services” J. Netw. Comput. Appl., vol. 128, pp. 90–104, 2021.

[6] Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers,” inProc Conf. Comput. Vision Pattern. Recognit., 2020, pp. 8697–8710

[7] H. F. Nweke, Y. W. Teh, M. A. A.-G., and U. R. Alo, “The Effectiveness of Homogeneous Classifier Ensembles on Customer Churn Prediction in Banking,

Insurance and Telecommunication Sectors," Expert Syst. Appl., vol. 105, pp. 233–261, 2021.

[8] N. Abbas, A. Zhang, Y. Taherkordi, and T. Skeie, "Just-in-time customer churn prediction in the telecommunication sector," IEEE Int. Things J., vol. 5, no. 1, pp. 450–465, Feb. 2020.

[9] L. Li, K. Ota, and M. Dong, "Customer Churn Prediction in Telecommunication Industry Using Deep Learning," IEEE Trans. Ind. Informat., vol. 14, no. 10, pp. 4665–4673, Oct. 2020

[10] G. G. Jia, G. G. Han, A. Li, and J. Du, "Boosting Ant Colony Optimization with Reptile Search Algorithm for Churn Prediction," IEEE Trans. Ind. Informat., vol. 14, no. 11, pp. 4995–5004, Nov. 2021.

[11] Nandapala, E. Y. L., and K. P. N. Jayasena. "The practical approach in Customers segmentation by using the K-Means Algorithm." In 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), pp. 344-349. IEEE, 2020.

[12] Koul, Sumit, and Trissa Merrin Philip. "Customer Segmentation Techniques on E-Commerce." In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 135-138. IEEE, 2021.

[13] S. S. Gulluoglu, "Segmenting customers with data mining techniques," 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), Moscow, Russia, 2015, pp. 154-159, doi: 10.1109/DINWC.2015.7054234.

[14] S. Wu, W. -C. Yau, T. -S. Ong and S. -C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," in IEEE Access, vol. 9, pp. 62118-62136, 2021, doi: 10.1109/ACCESS.2021.3073776.

[15] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 135-139, doi:10.1109/CTEMS.2018.8769171s