EXPLORING CUSTOMER SEGMENTATION IN THE IT SECTOR USING DATA MINING TECHNIQUES

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

[1]Associate professor, [3,4,5]Student, Department of CSE, Erode Sengunthar Engineering College, Perundurai, Erode.

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

ABSTRACT

With its wide range of clients, the IT sector consistently produces large amounts of data. Cost reasons are one reason why business professionals emphasise the importance of keeping current clients above attracting new ones. Understanding behavioural trends and client attrition issues is therefore crucial. In the present work, a churn prediction model using classification and clustering techniques specifically tailored for the IT industry is proposed. Information gain and correlation attribute ranking filters are included in feature selection.Effective algorithms for identifying churn cases include Random Forest (RF), Naive Bayes, Multilayer Perceptron (MLP), and Logistic Regression (LR). Important churn variables are also identified by the model, which makes it possible to create retention plans that work. After classifying customers, the algorithm uses cosine similarity to segment them and provide tailored retention offers.

Keywords: Churn, Classifications, Clustering, Correlation attribute, Information gain,Segment.

1.INTRODUCTION

Within the ever-changing Information Technology (IT) sector, businesses are constantly looking for new and creative ways to get a competitive advantage and provide customised solutions for their customers. Customer segmentation is an important tactic in this endeavour. IT companies may successfully classify their heterogeneous customer base into discrete groups, each with specific qualities and requirements, by utilising data mining technologies. With this strategy, businesses may better understand consumer preferences, allocate resources more effectively, and create customised client acquisition and retention plans.

1.1 ASSOCIATION RULE MINING

Association rule mining, a reliable data mining technique, is extensively used in a number of sectors, including retail, market basket research, and recommendation systems. This method explores databases to find meaningful and fascinating relationships between components. Association rule mining helps companies make better decisions and run more efficiently by exposing patterns and dependencies in data. The process

basically consists of finding shared item sets and creating rules that explain why items appear together in transactional or category datasets. These principles are essential tools for gleaning insightful information from complex and sizable datasets, revealing underlying relationships that may be difficult to notice. Finding correlations or connections between components that occur together more frequently than would be predicted by chance is the main goal of association rule mining.

1.2 CUSTOMER SEGMENTATION

Segmentation of customers is a strategic technique used by businesses and organisations to divide their customer base into discrete groups according to shared traits, habits, and preferences. By using this method, companies can obtain a better understanding of their varied clientele, which makes it easier to create products, services, and marketing plans that are specifically suited to each group's requirements. Acknowledging each customer's uniqueness improves productivity, client happiness, and overall market competitiveness. Customer segmentation helps organisations handle resources more effectively and build more customised, strong client relationships. It also acts as a road map for targeted marketing initiatives. Customer segmentation is essential for understanding complex customer patterns and trends in today's data-driven world. For segmentation reasons, a variety of factors are used, including location, historical purchasing history, demographics, and psychographic information. Businesses may improve customer service programmes, hone product offers, and create more persuasive marketing messages by utilising these data-driven categories. This will ultimately boost brand loyalty and increase customer retention rates.

1.3 MARKET ANALYSIS

A fundamental aspect of sound business decision-making involves conducting market studies, which entail comprehensive analysis of specific markets or industries to grasp their potential, trends, and dynamics. Employing this analytical technique, businesses can glean crucial insights into aspects such as market size, growth prospects, competitive landscape, and consumer behavior. These insights are instrumental in formulating strategic plans, making informed investments, and adapting to the ever-evolving market conditions. Market analysis is indispensable for companies striving to maintain competitiveness and relevance in the global marketplace, particularly amidst rapid technological advancements and evolving consumer preferences. Market analysis covers a wide range of activities, including competitive evaluation, risk and opportunity identification, and data collection and analysis on consumer preferences and market demographics. It is the starting point for important decisions about product development, pricing strategies, market entry, and marketing campaigns.

2. LITERATURE REVIEW

[1] Targeted marketing strategy is a hot topic that has drawn a lot of interest from academics and industry, as suggested by Fahed Joseph et al. in this research. A popular method for examining the diversity of consumer purchasing behavior and profitability is market segmentation. It

is noteworthy that traditional models of market segmentation used in the retail sector are mostly descriptive in nature, do not provide adequate market insights, and frequently do not identify small enough categories. In order to process large amounts of data, this study also makes use of the dynamics present in the Hadoop distributed file system. Expectation-Maximization (EM) and K-Means++ clustering algorithms were used in three separate market segmentation studies utilizing modified best fit regression. The results were evaluated using cluster quality evaluation. The study's findings are as follows: (i) each consumer Lifetime Value (CLTV) segment's insight into consumer buying behavior; (ii) the clustering algorithm's performance in creating precise market segmentation. Based on the data, the average customer lifetime was found to be just two years, with a 52% churn rate. As a result, a marketing plan was created based on these findings and applied to department store sales.

[2] E. Ernawati et al. suggested that the process of extracting knowledge from data is known as data mining (DM). Businesses can identify their target market and develop a marketing strategy with the use of the data obtained from customer behaviour segmentation. The Regency Frequency Monetary (RFM) model is the behaviour segmentation model that is most frequently employed. In several customer-segmentation studies across multiple application domains, the RFM model collaborates with DM. With so many DM approaches at one's disposal, picking the appropriate ones could lead to the discovery of interesting hidden patterns in client segments. The aim of this study is to establish a framework for consumer segmentation by analysing and synthesising DM strategies that work with the RFM model. This study employs a comprehensive review of the literature spanning the years 2015–2020. Grouping and visualisation are two of the seven DM strategies that have been studied that are used the most frequently. Because of the expanded visualisation function and the need to consider customers' geo-demographic data in the analysis, this study suggests a new framework for combining DM approaches with the RFM based segmentation in the Geographic Information Systems (GIS) environment.

[3] SHULI WU et al. suggest that hiring new customers is no longer a smart business strategy in the telco sector because retaining existing ones is less expensive than getting new ones. Churn management becomes essential in the telecom industry. Since there isn't much research combining customer segmentation and churn prediction, this study aims to propose an integrated customer analytics approach for churn management. Factor analysis, churn prediction, customer segmentation, customer behaviour analytics, exploratory data analysis (EDA), and data pre-processing are the six components that make up the framework. This technology integrates the customer segmentation process with churn prediction to provide telecom operators with a thorough churn analysis and assist them in better managing customer attrition. Three datasets are used in the testing with six machine learning classifiers. Initially, a number of machine learning classifiers are employed to predict the customers' likelihood of leaving. The training set is exposed to the Synthetic Minority

Oversampling Technique (SMOTE) in order to overcome the problems associated with imbalanced datasets. The 10-fold cross-validation is used to assess the models.

[4] Saumendra Das et al suggest that customers are growing in awareness, education, and social engagement. To satisfy their wants, they modify their routines and tastes. The concept of segmenting customers is to divide heterogeneity into homogeneous forms. Customer segmentation is an essential part of marketing since it helps companies manage massive amounts of customer data in an organised manner and build relationships with customers. Understanding the client's hidden knowledge is a deft use of computational analysis, which allows for the customisation of exact data to the client's interests and preferences. This type of computational analysis is known as "data mining." This essay methodically looked at consumer segmentation utilising data mining techniques. It is a structured examination of various supervised and unsupervised segmentation-related data mining techniques.

[5] Angel Martín et al. assert that global navigation satellite system (GNSS)-based location and navigation services are essential for real-time high-precision positioning applications in important economic sectors like transportation, civil engineering, precision agriculture, and mapping. Globally, the use of GNSS networks for real-time navigation has increased significantly since the 1990s and has surpassed initial predictions. Therefore, by monitoring the changes in GNSS network users, market segment patterns can be investigated. To implement this concept, a significant volume of navigation data needs to be analysed over a number of years, and clients need to be closely watched. This project intends to collect statistics and analysis with the primary purpose of managing large-scale GNSS user connections efficiently. The results demonstrate the changing needs over time as well as the characteristics of users in different market segments.

3.EXISTING SYSTEM

Retail marketers are always looking for methods to make their campaigns more effective. Targeting consumers with incentives that are likely to entice them to return and increase their spending and time spent there is one tactic. One technique used for this is demographic market segmentation. It entails grouping the larger market according to preset standards. Factors including age, gender, marital status, occupation, education, and income are frequently used in demographic segmentation. A Turkish grocery chain was the subject of an example case study to demonstrate how segmentation ideas are applied. This case study aims to identify the relationships between products and purchasing habits, and the advertising strategies for products and consumer profiles are guided by expected sales numbers.

4. PROPOSED SYSTEM

The IT industry's suggested churn prediction model uses data-driven methodology to pinpoint clients who are leaving and the reasons behind their departure. To properly accomplish its goals, it incorporates approaches for clustering and classification. To ensure that the most pertinent attributes are chosen for churn prediction, the model

first employs feature selection algorithms that leverage correlation attribute ranking filters and information gain. After that, it classifies churn customer data using four robust classification algorithms: Random Forest, Naive Bayes, Multilayer Perceptron, and Logistic Regression. After categorising the churning customer data, the model uses cosine similarity to divide it into clusters. This makes it easier to create personalised retention incentives, like deals or promotions that are specifically catered to the clusters.

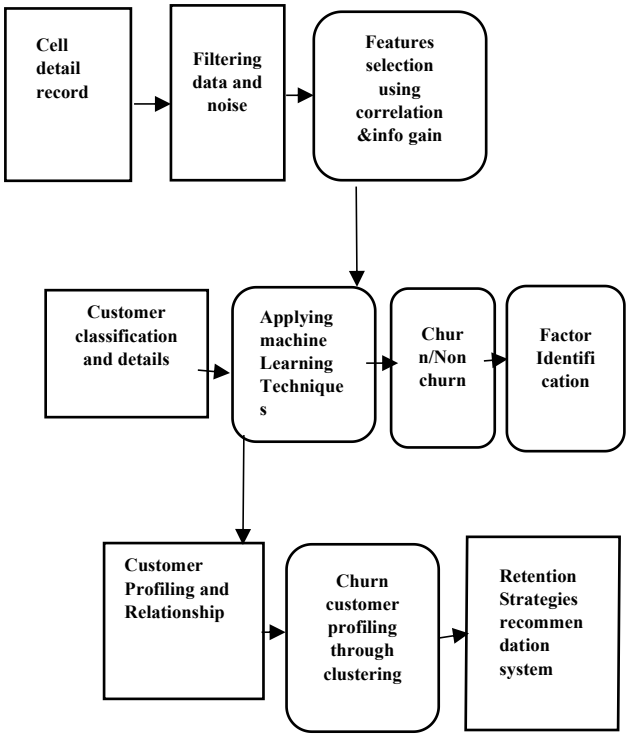

Figure 1 Block Diagram

To put it simply, the model pinpoints critical churn factors that are necessary to comprehend the root causes of churn. CRM teams can use this information to improve their retention tactics and create marketing campaigns that are more successful. All things considered, IT companies can use the suggested churn prediction model as a useful tool to lower customer attrition, boosting customer happiness and loyalty.

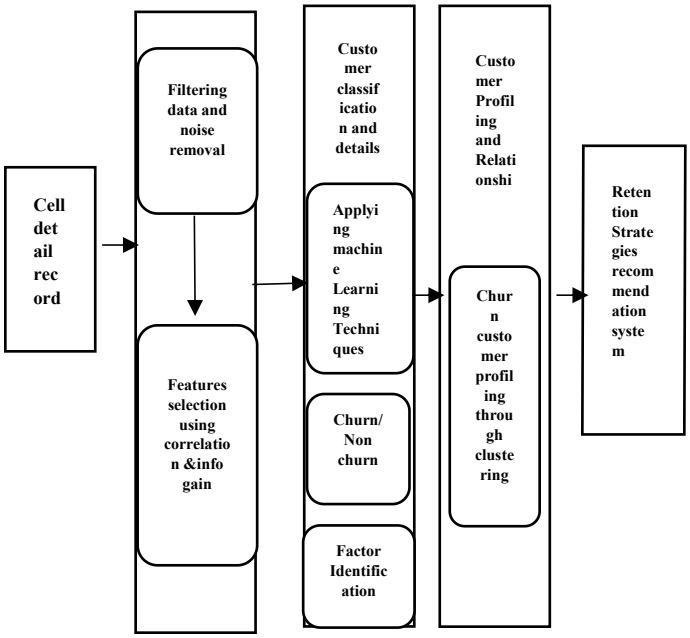

Figure 2 FLOW DIAGRAM

4. MODULE DESCRIPTIONS

4.1 LOAD DATA

Obtaining and importing pertinent data into the project's data analysis environment is the first step in the process. Loading data is an essential stage since it supplies the raw material needed for further analysis. Preparing and organising the data appropriately is crucial for successful processing and analysis.

4.2 DATA PRE PROCESSING

The loaded data goes through a number of procedures at this point in order to get it ready for analysis. The data pre-processing duties include things like controlling outliers that could skew the analysis, removing duplicates, and cleaning the data to handle missing or inaccurate numbers. To make sure the data is consistent and

appropriate for machine learning algorithms, methods like data transformation and standardisation are also used.

4.3 FEATURE SELECTION

The critical phase of feature selection involves determining the characteristics or factors that are most important in forecasting client turnover. This approach is guided by a number of strategies, including domain expertise, correlation analysis, and knowledge development. The dimensionality of the data is decreased by choosing a subset of the most important attributes, which may result in more precise and accurate churn prediction models.

4.4 TRAINING AND TESTING

The project moves on to the training and testing phase after the data has been cleaned and pertinent features have been found. Machine learning models are built and trained using historical data as a basis. A portion of the data is usually set aside for model testing and assessment of predictive power. A range of classification approaches are utilised to assess how well they can differentiate between churners and non-churners, including Random Forest, Naïve Bayes, Multilayer Perceptron, and Logistic Regression.

4.5 PERFORMANCE AND EVALUATION

The churn prediction models' performance is evaluated in the last phase. A number of assessment metrics, such as recall, accuracy, precision, and F1-score, are employed to determine how well the models detect churn. These measures demonstrate the models' efficacy and possible real-world uses. A thorough performance assessment stage is essential in order to make well-informed decisions about marketing campaigns and client retention strategies based on the model's predictions.

5.ALGORITHM DETAILS

The suggested churn prediction methodology in the IT sector uses a combination of clustering and classification techniques in two phases to discover and understand churning clients using a data-driven approach. The model first uses information gain and correlation attribute ranking filters to perform feature selection, ensuring that the most relevant features are used for churn prediction. It then uses four strong classification algorithms to categorise client data according to their probability of churning: Random Forest, Naive Bayes, Multilayer Perceptron, and Logistic Regression. These carefully chosen classification algorithms provide a range of perspectives on consumer behaviour. Following categorization, the model clusters the churning customer data using cosine similarity, which makes it easier to create unique customer segments. Because of this segmentation, specialised retention methods that are catered to the unique traits of each identified segment can be developed, such as customised offers or promotions. By using cosine similarity, the model becomes more adept at identifying minute patterns and similarities among customers who are about to leave, which in turn makes retention campaigns more targeted and efficient.

6.RESULT ANALYSIS

The performance of these methods inside their respective domains is demonstrated by the accuracy scores given for the

Apriori algorithm and the ensemble of Random Forest (RF), Naive Bayes (NB), Multilayer Perceptron (MLP), and Logistic Regression (LR). With an accuracy rate of 70%, the Apriori algorithm—which is commonly used in association rule mining—shows that it can find patterns and associations in the dataset. On the other hand, the group of classification algorithms (RF, NB, MLP, and LR) as a whole obtains an accuracy rate of 85%, which suggests that they are good at forecasting results for the provided dataset, which may be associated with a classification task such as customer churn prediction. The increased accuracy indicates that using a variety of different classification strategies improves overall predictive power when compared to the Apriori algorithm alone, which makes it a better option for the particular task at hand—possibly in the context of predicting customer churn in the IT industry.

Algorithm	accuracy
Apriori	70
Rf,Nb,Mlp,Lr	85

Figure 3 comparison Table

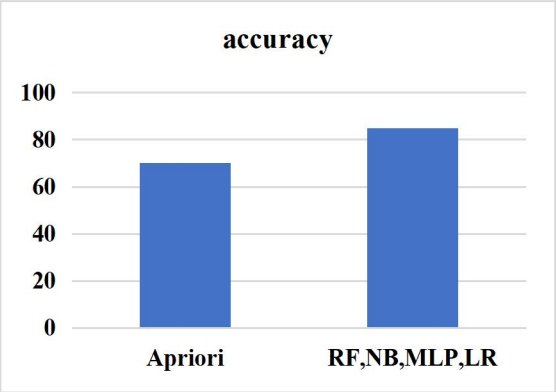

Figure 4 comparison Graph

5. CONCLUSION

In conclusion the recommended churn prediction model presents IT companies with a valuable tool to reduce customer attrition, enhance customer satisfaction and loyalty, and increase revenue. The model is scalable, actionable, comprehensive, and precise. It takes into account various factors contributing to customer attrition and provides IT businesses with actionable insights for crafting effective retention campaigns. Moreover, the utilization and implementation of the model are straightforward. It does not necessitate specialized expertise for operation and can easily integrate with existing CRM systems.

7. FUTURE WORK

Expanding the model's data sources represents a potential avenue for future development. Incorporating additional data from customer support interactions, social media activities, and surveys could enhance the model's predictive capabilities and provide a more holistic understanding of individual consumers. Additionally, exploring new churn prediction algorithms and methodologies is a promising area for further research. For example, deep learning systems have shown impressive performance in various classification tasks, including churn prediction. Examining the feasibility of integrating deep learning techniques to improve the suggested churn prediction model would be an intriguing area of investigation.

8. REFERENCES

[1]Fahed Yoseph and Markku Heikkila," The Impact of Big Data Market Segmentation Using Data Mining and

Clustering Techniques",CISCO, Cisco Global Cloud Index: Forecast and Methodology, 2020

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

[12] S. Koul and T. M. Philip, "Customer Segmentation Techniques on E-Commerce," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)

[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