Customer Churn Prediction Using Synthetic Minority Oversampling Technique

Aishwarya H M
dept. Artificial Intelligence & Machine Learning
Global Academy of Technology
Bangalore, India
aishugowda3007@gmail.com

Bindhiya T
dept. Artificial Intelligence & Machine Learning
Global Academy of Technology
Bangalore, India
bindhiyathyagaraj@gmail.com

S Tanisha
dept. Artificial Intelligence & Machine Learning
Global Academy of Technology
Bangalore, India
tanishasunke@gmail.com

Soundarya B
dept. Artificial Intelligence & Machine Learning
Global Academy of Technology
Bangalore, India
soundaryakash12@gmail.com

C Christlin Shanuja
dept. Artificial Intelligence & Machine Learning
Global Academy of Technology
Bangalore, India
christlinsha@gmail.com

2023 4th International Conference on Communication, Computing and Industry 6.0 (C216) | 979-8-3503-2732-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/C2I659362.2023.10430989

***Abstract*— Customer churn prediction is the process of using data and analytics to forecast that customers are likely to stop using a product or service. It helps businesses proactively identify and retain at-risk customers by implementing targeted retention strategies. Business analysts and CRM (Customer Relationship Management) analysts should comprehend the effects behind customer churn and analyze the behaviour trends of existing churn customers from the data. From telecom sector, gaining new consumers is not considered as a smart move because it costs far less to keep existing ones. This approach combines customer segmentation and churn prediction to give telecom operators a comprehensive customer churn analysis. It helps in businesses reduce costs, retain customers through personalized interventions, and gain a competitive edge by leveraging data-driven insights to improve overall profitability and customer satisfaction. Four machine learning (ML) classifiers are employed in the experiments. First, these ML classifiers are used to predict the customer's churn state. To focus on the imbalanced datasets issues, we apply the proposed Synthetic Minority Oversampling Technique (SMOTE). The best method is Gradient Boosting Classifier, which has achieved an accuracy of 95.13 percent, based on the experimental investigation.**

Keywords—Customer churn prediction, Customer Relationship Management, Machine Learning, Synthetic Minority Oversampling Technique, Gradient Boost

I. INTRODUCTION

Customer churn is the term used to describe the phenomenon of customers or subscribers discontinuing a business with a firm or service. It is a crucial business application of ML that helps industries identify and retain valuable customers. Companies typically distinguish between involuntary and voluntary churn. A company may experience involuntary churn if their clients relocate to a distant location, move into a long-term care facility, or pass away. Voluntary churn occurs when a consumer chooses to migrate to another business or service provider. In the majority of applications, analytical models do not include involuntary churn factors.

The origination of studies on the proposed work began from Customer Relation Management (CRM). The CRM is a corporate management technique that is to boost the efficiency of an organization and customer value functions in the domains of supply chain management, retail, marketing, sales, and customer support [1].

Reactive and proactive approaches are the two categories of customer churn management strategies [2]. Majority of recent studies, for instance [3], considered the proactive approach to be a more effective approach. Predicting and preventing churn is essential for businesses to maintain revenue, customer loyalty, and overall growth. These providers frequently employ customer attrition analysis and customer attrition rates as one of their primary business metrics as the retaining cost with an existing customer is significantly less than acquiring a new one.

In the Telecom industry, the two main elements of customer analytics include customer segmentation and churn prediction. Since it is less expensive to identify and keep current customers than to find new ones, the telecom companies have developed tactics to do just that [4]. This is because maintaining existing customers can cost nearly five or six times less than as much as advertising, concessions and labour. It is difficult to identify the cause of churn and its frequency. Several problems with customer development, involving those related to service quality, network coverage, load failures, billing, charges, technology, etc., are mostly caused due to quality component. Customers may assess benefits and service quality with another suitable service provider using the above-mentioned service quality factors [5].

We present a proposed customer churn prediction model that utilizes various ML techniques. The classifier performance is based on dataset efficiency that has been validated using a real-world dataset named Telco Customer Churn. The proposed work has been evaluated using the information retrieval measures with Precision, Recall, F-measure and the churn prediction model accuracy is computed. The studies employ the following algorithms: Gradient Boosting Classifier, Decision Tree Classifier, Random Forest Classifier, and Logistic Regression. The project is to investigate existing machine learning techniques, propose a model for anticipating client attrition,

identify factors that cause churning, and provide retention tactics. Based on the experiments, we observed that our suggested model performed better than the competition in terms of accurately classifying churners. Using SMOTE, we found that the Gradient Boosting Classifier outperformed other machine learning methods with regard to accuracy.

II. LITERATURE SURVEY

Churn research is typically experimented out to enhance business results. Therefore, the churn time is described as a period that can regain back customer's trust in the number of churn prediction problems. If the time range during which a client entirely leaves is chosen, the period for churn definition rapidly grows and offers no business benefit because it is thought to be impossible to change the minds of consumers who want to leave. The contracts stated above come very near to completely excluding customers from a service. Because of this, the probabilistic approach is nowadays used by most log-based churn prediction challenges to identify whether or not customers are churning and to offer incentives to keep them using the service.

Majority of early studies on churn prediction in the telco industry concentrated on a limited set of ML classifiers. They ran few experiments on various feature sets using the six prediction methods (AdaBoost, Multiple Layer Perceptron, Naive Bayes, Decision Tree, Random Forest, and Logistic Regression) [6] to make observation and choose the best predictive model. With accuracy, precision, Recall and F1-Score as 93.60%, 74.63% 80.71%, and 77.20% respectively [6], Random Forest fared the best across all measures. It is clear that Decision Tree & Random Forest perform more accurately, with Random Forest achieving the greatest AUC (area under the ROC curve) of 91.40%.

In the study of [6], more than 100 classifiers efficiency was evaluated in connection to the churn prediction problem faced by the telecom sector. According to existing results, Regularized Random Forest exceeds all other classifiers with an 73.04% accuracy score while Bagging Random Forest topped them all with over an AUC of 67.20%.

To more accurately identify high-value clients, suggested an innovative customer segmentation strategy based on the customer life cycle. Direct value was divided into categories, encompassing historical, long-term, current, and indirect values. These five component's calculations result in five models. Experts assigned weights to these five factors, and ranking was used to determine the high-value clients.

The author [7] developed a consumer segmentation approach with K-means clustering and the commercially automated technology KXEN. Small business clients were parted into two categories from on their values and behaviours. Customers were divided into five groups according to their behaviour and six groups according to their value, according to the crossing matrix that was displayed. Through the analysis of each segment's client characteristics, marketing analysts were able to observer the needs of their target market and suggest companies or packages that they would like to use.

In the study of [8], customer turnover is predicted using Logistic regression and Decision trees. According to their discovery, Decision Tree performed better on their datasets than Logistic Regression. By utilizing a Decision Tree classifier, they achieved a maximum 99.67% accuracy on their sizable dataset. Another study [9], utilized KNN (K-Nearest Neighbor), Random Forest, and XGBoost classifiers to estimate the attrition of customers. They found that XGBoost gained the highest accuracy (79.8%), along with the highest AUC (58.2%). The study [10] employed Random Forest, AdaBoost, C4.5, KNN, Naive Bayes, and ANN to predict client turnover. Using the Synthetic Minority Over Sampling Technique (SMOTE) the dataset's instances were balanced efficiently.

In order to find the most effective model, many parameter combinations of various techniques were investigated. KNN, C4.5, and Random Forest classifiers did well in their on AUC value. Based on results, Random Forest surpass top, scoring 91.10%. They suggested that the older dataset's data tends to a negative impact on the outcomes of their experiments. Random Forest and Naive Bayes were also used by [11]. Their findings tells that Random Forest classifier, with an 71.99% accuracy of outperformed Naive Bayes.

III. RESEARCH MOTIVATION AND CONTRIBUTIONS

The cost of recruiting new consumers is five to ten times more than the cost of keeping current ones, and new customers frequently leave more frequently than current ones.

Furthermore, there are presently more telecom providers in the same areas, and they are all constantly improving their own products and services. Users have too many options in this situation. As the shifting cost of telecom users declines, telecom operators find it more difficult to hold onto their current customer base. The high service standards that telecom customers have led to significant marketing expenses for telecom operators. Extensive telecom databases are the phases of the daily information generation and vast customer base. The fact that telecom operator customers do not simply switch over a limited period of time, however, poses a serious challenge for churn prediction process. Majority will decide against switching providers and tends to use their current operators' services. This circumstance leads to an imbalance in the telecom sector dataset, that has a significant impact on churn prediction [12, 13].

The proposed work SMOTE is used on the training data, and several metrics are corelated and explained. Although the operator can only identify the churn customers, if just churn prediction is done, it is unable to fully comprehend the causes behind it. If customer segmentation alone is used, the operator will only be able to understand the features of the various clusters' customers, that will prevent it from concentrating on churning customers and causing it to react differently to those clusters.

Research on telecom consumers must be restricted to figuring out which customers are most probably to leave. Finding the causes of client turnover is increasingly crucial. Furthermore, churn factors are included in the models to help marketers better in understanding the churn causes.

IV. METHODOLOGY

A. Implementation

The system uses raw data from many sources, such as client transactions, interactions, and historical data. Data sources, ETL

(Extract, Transform, Load) processes, and data warehouses are all examples of data sources.

To prepare it for model training, raw data is cleaned, converted, and pre-processed. Data cleaning tools, feature engineering, normalization, and scaling are all available. Creating meaningful features from raw data to improve the model's predicting performance. Methods for feature selection, domain knowledge, and statistical approaches. Using historical data, build and train ML models to predict customer attrition. Model training environment, hyperparameter adjustment, and ML techniques (e.g., logistic regression, decision trees, and random forests). Evaluating the trained model's performance using evaluation measures to confirm its correctness and generalizability. Select a machine learning model that works well for problems involving binary classification. Neural networks, support vector machines (SVM), random forests, decision trees, and logistic regression are popular methods for predicting customer attrition. Divide the data into sets for testing and training. Utilizing the training data, train the selected model and test model parameters to maximize performance. Through data analysis, the model discovers linkages and patterns that can be leveraged to forecast future events. To evaluate the model on the testing set, use relevant performance metrics such as area under the receiver operating characteristic curve (AUC-ROC), recall, accuracy, precision, and F1 score. These measurements give us information about how well the model detects churn.

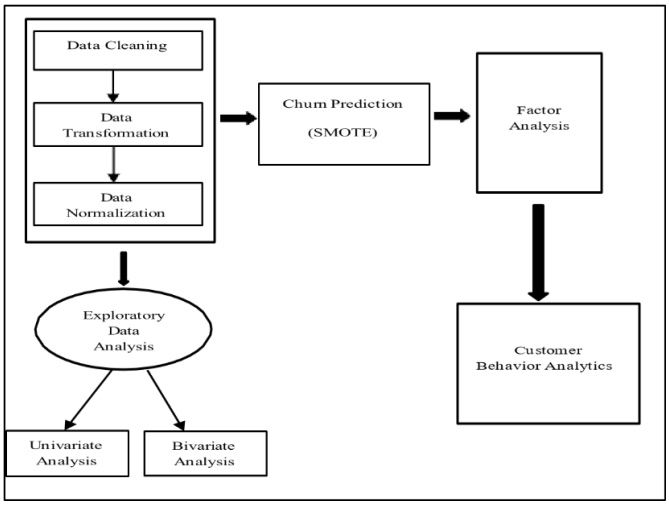

Fig. 1. Architectural Diagram of Customer Churn Prediction.

New customers often have higher attrition rates than current ones, and acquiring new clients comes at a five-to ten-fold higher expense than keeping current ones. The implementation of a retention strategy for telco operator's approach can increase earnings by restoring the confidence of clients that have previously used the products and services. However, the number of telco providers in the same these days, and each operator is continually improving its own products and services. In this situation, telco clients offer too many options. As telco customer costs fluctuate diminishes, telco companies find it harder and harder to hold onto customers.

Later, research on churn was actively done in both the telecommunications and gambling industries. Due to intense rivalry, many services. experience a rapid cycle of customer acquisition and attrition. The Customer Acquisition Cost (CAC) tends to rise, however, if a single service is provided for an extended period of time [14]. The need for churn prediction and prevention technologies grows as the CAC expands. The author [15] used the survival analysis to analyze mobile games and determined the churn rate in a manner akin to how financial services forecast churn. Due of the abundance of log data, the gaming industry actively uses machine learning approaches to study churn.

Churn research is typically made to enhance business results. If the time range during which a client entirely leaves is chosen, the period for churn definition exponentially grows and offers no business benefit because it is thought to be impossible to change the minds of consumers who want to leave. There are numerous ML and data mining technologies available that are used to calculate such data as a outcome of most recent developments in the Big Data field. These methods examine the data to determine the causes of client churn. Customer Churn Management (CRM) can use these strategies to increase their revenue.

Additionally, it can be utilized to create retention plans that will lower the proportion of consumers who will leave. By accurately detecting customers' wants through the use of data mining tools, the CRM can help a business meet its goal of client retention. Data mining is the technique of determining customer turnover behaviour based on patterns gleaned from the data. In the pre-processing stage, firstly, data cleansing, data transformation, and data normalization are carried out.

The subsequent phase is exploratory data analysis (EDA), which includes univariate and bivariate analysis. The purpose of employing EDA is to help us better understand the data before providing each feature to machine learning models, which will improve the modelling process. Second, four machine learning classifiers - Logistic Regression, Decision Tree, and Random Forest, Gradient Boosting - are employed to predict customer attrition. Thirdly, factor analysis is carried out by fine-tuning the gradient boosting technique in order to identify certain significant characteristics underlying the churning.

Finally, consumer behaviour analytics are carried out based on the results of churn prediction. The churn uncertainty is represented visually. All different machine learning classifiers are evaluated and compared.

B. Dataset

A customer is represented by each row, and the properties of each client are described in each column's metadata. Customers who have departed during the last month are indicated in the churn column. The phone, multiple lines, internet, gadget protection, online security, online backup, tech support, and TV and movie streaming are among the services each user has signed up for. Information on a customer's account, such as the length of the contract, the mode of payment, paperless invoicing, the amount due each month, and the overall charges. Gender, age

range, and presence or absence of partners and dependents are examples of customer demographics.

The dataset contains 7044 rows and 21 columns, in which all attributes are numerical. Since the test dataset lacks the customer's churn labels, only the train datasets is used for the experiment. Only 16.2% customers are senior citizens but remaining 83.8% customers are young people [1, 11].

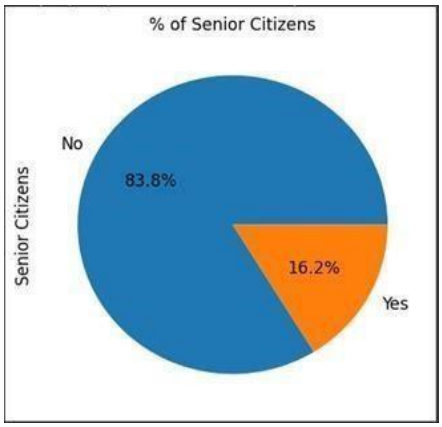

Fig. 2. Count of Senior Citizens.

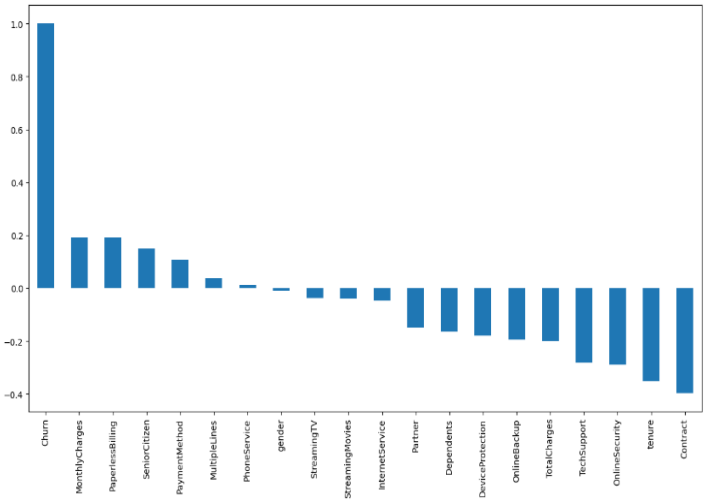

Fig. 3. Correlation of churn with other attributes.

V. RESULTS

The results of utilizing SMOTE Logistic regression gained an accuracy rating of 89.04%. The accuracy rates for Decision Tree and Random Forest are 91.81% and 92.03%, respectively. The accuracy rating for Gradient Boost was 93.80%. After tuning by altering the criteria, learning rate, loss, maximum depth, maximum leaf nodes, minimum samples leaf, minimum samples split, and N estimators, it attained an accuracy score of 95.13 percent. This time, churn control is improved since more customers who genuinely leave are identified than in previous studies without SMOTE.

Without employing SMOTE, the results Logistic Regression had an accuracy rating of 81.19%. The accuracy rates for Decision Tree and Random Forest are 80.83% and 81.61%, respectively. As a result of the anticipated outcomes, churn management is ineffective since it cannot detect many consumers who actually churn.

TABLE I. ATTRIBUTES OF DATASET FOR SEGMENTATION

#	Attributes	#	Attributes	#	Attributes
1	customerID	**8**	MultipleLines	**15**	StreamingMovies
2	gender	**9**	InternetService	**16**	Contract
3	SeniorCitizen	**10**	OnlineSecurity	**17**	PaperlessBilling
4	Partner	**11**	OnlineBackup	**18**	PaymentMethod
5	Dependents	**12**	DeviceProtection	**19**	MonthlyCharges
6	tenure	**13**	TechSupport	**20**	TotalCharges
7	PhoneService	**14**	StreamingTV	**21**	Churn

C. *Performance Metrics*

The proposed work uses accuracy, precision, recall, and F1-score to evaluate the proposed churn prediction model. Accuracy is measured as the correctly classified points by a total no of points on the test set.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Precision can be described as the proportion of accurately predicted positive observations to total expected positives. It stresses the precision of favourable forecasts.

$$Precision = \frac{True\ Positive}{(True\ Positive + Fasle\ Positive)} \qquad (2)$$

TABLE 2. RESULTS ANALYSIS WITH SMOTE FOR DATASET

Methods	Accuracy	Precision		Recall		F1-score	
		Class 0	Class1	Class0	Class1	Class0	Class1
Logistic Regression	89.04	88.00	90.00	89.00	89.00	88.00	90.00
Random Forest	92.03	92.00	92.00	91.00	93.00	92.00	92.00
Decision Tree	91.81	92.00	92.00	91.00	92.00	91.00	92.00
Gradient Boost	95.13	94.00	96.00	95.00	95.00	95.00	95.00

TABLE 3. RESULTS ANALYSIS WITHOUT SMOTE FOR DATASET

Methods	Accuracy	Precision		Recall		F1-score	
		Class 0	Class1	Class0	Class1	Class0	Class1
Logistic Regression	81.19	90.00	55.00	85.00	66.00	88.00	60.00
Random Forest	81.61	93.00	48.00	84.00	71.00	88.00	57.00
Decision Tree	80.83	93.00	46.00	83.00	70.00	88.00	55.00

Applying SMOTE can improve predictive performance in situations where class imbalance has a detrimental impact on model performance. More balanced datasets tend to provide models that generalize better to new data. Training times may increase as a result of SMOTE's increase in dataset instances, particularly for algorithms that are sensitive to the size of the training set. Although SMOTE aids in balancing the distribution of classes, noise may also be introduced into the dataset. If the artificial examples don't accurately reflect the distribution of the underlying data, overfitting may happen.

VI. CONCLUSION

Businesses can gain accurate insight into the selection of their consumers with machine learning-based customer churn analysis. The establishment of a churn prediction model, which helps telecom operators in predicting which customers are most likely to depart and improving services accordingly, is the main contribution of this research. Applying the SMOTE Gradient Boost algorithm, which has an accuracy rating of 95.13%, produced the best results. Good churn modelling performance was also demonstrated by other models, such as the Random Forest Classifier and Decision Tree Classifier. Better model evaluation criteria and imbalanced datasets were addressed with the aid of SMOTE. These business models support telecom services in becoming lucrative.

Future research will focus on building other churn prediction models, such as SMOTE variants and deep learning models, and trials will be run to develop models that are more accurate. Further research into ROC analysis may be done in an effort to develop a more long-lasting churn prediction model.

It also might focus on designing algorithms, using numerous data sources, and enhancing model interpretability. These innovations promise to provide businesses more options to keep customers, boosting their competitiveness and profitability.

REFERENCES

[1] A. Parvatiyar and J. N. Sheth, ‘‘Customer relationship management: Emerging practice, process, and discipline,’’ J. Econ. Social Res., vol. 3, no. 2, 2001.

[2] C.-F. Tsai and Y.-H. Lu, ‘‘Customer churn prediction by hybrid neural networks,’’ Expert Syst. Appl., vol. 36, no. 10, pp. 12547–12553, Dec. 2009.

[3] G. F. Retana, C. Forman, and D. J. Wu, ‘‘Proactive customer education, customer retention, and demand for technology support: Evidence from a field experiment,’’ Manuf. Service Oper. Manage., vol. 18, no. 1, pp. 34–50, Feb. 2016.

[4] A. Idris and A. Khan, ‘‘Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers,’’ in Proc. 15th Int. Multitopic Conf., Dec. 2012, pp. 23–27.

[5] W. Verbeke, D. Martens, C. Mues, and B. Baesens, ‘‘Building comprehensible customer churn prediction models with advanced rule induction techniques,’’ Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, Mar. 2011.

[6] Shuli Wu, Wei-Chuen Yau, (Member, IEEE), Thian-Song Ong, (Senior Member, IEEE), And Siew-Chin Chong, (Senior Member, IEEE) S. Wu et al, “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business,” 2021.

[7] L. Ye, C. Qiuru, X. Haixu, L. Yijun, and Z. Guangping, ‘‘Customer segmentation for telecom with the k-means clustering method,’’ Inf. Technol. J., vol. 12, no. 3, p. 409, 2013.

[8] K. Dahiya and S. Bhatia, ‘‘Customer churn analysis in telecom industry,’’ in Proc. 4th Int. Conf. Rel., Infocom Technol. Optim. (ICRITO), Sep. 2015, pp. 1–6

[9] J. Pamina, B. Raja, S. SathyaBama, S. Soundarya, M. S. Sruthi, S. Kiruthika, V. J. Aiswaryadevi, and G. Priyanka, ‘‘An effective classifier for predicting churn in telecommunication,’’ J. Adv. Res. Dyn. Control Syst., vol. 11, no. 10, pp. 221–229, Jun. 2019.

[10] G. Esteves and J. Mendes-Moreira, ‘‘Churn prediction in the telecom business,’’ in Proc. 11th Int. Conf. Digit. Inf. Manage. (ICDIM), Sep. 2016, pp. 254–259.

[11] E. Lee, B. Kim, S. Kang, B. Kang, Y. Jang, and H. K. Kim, ‘‘Profit optimizing churn prediction for long-term loyal customers in online games,’’ IEEE Trans. Games, vol. 12, no. 1, pp. 41–53, Mar. 2020

[12] T. J. Gerpott, W. Rams, and A. Schindler, ‘‘Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market,’’ Telecommun. Policy, vol. 25, no. 4, pp. 249–269, May 2001.

[13] G. Olle, ‘‘A hybrid churn prediction model in mobile telecommunication industry,’’ Int. J. e-Educ., e-Bus., e-Manage. e-Learn., vol. 4, no. 1, p. 55, Feb. 2014.

[14] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, ‘‘Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry,’’ IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 690– 696, May 2000

[15] G.-E. Xia and W.-D. Jin, ‘‘Model of customer churn prediction on support vector machine,’’ Syst. Eng. Theory Pract., vol. 28, no. 1, pp. 71–77, Jan. 2008.