Volume 13, Issue 2
April 2024

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Customer Churn Prediction for the Telecom Industry

[1]Ashutosh Dashrath Kashid, [2]Akash Shankar Jadhav, [3]Vaibhav E. Narawade

[1,2] Department of Computer Engineering, Ramrao Adik Institute of Technology, Nerul, India.

[3]Professor, Department of Computer Engineering, Ramrao Adik Institute of Technology, Nerul, India.
E-mail: [1]ashutoshkashid5@gmail.com , [2]jadhavakashshankar@gmail.com, [3]vaibhav.narawade@rait.ac.in

ABSTRACT

Customer churn, refers to the rate at which customers are lost by a business. In the telecommunications industry, customer churn is a critical metric used to anticipate the customers who may switch from one telecom service provider to another. Churn prediction in the telecom industry is about using data & predictive analytics to foresee which customers might leave, allowing companies to implement strategies to keep them engaged and maintain revenue, while also avoiding the costly process of acquiring new customers. This article is expected to provide valuable insights for researchers and data analysts within the telecommunications sector. It aims to assist them in making informed decisions regarding the selection of optimal and suitable methodologies, as well as in the development of advanced and innovative churn prediction models in the future.

Keywords: *Churn, Telecom Industry, Predictive analysis, Methodology, Telecommunication, Customer churn prediction, Data Analysis, Research, Churn prediction models.*

1 INTRODUCTION

The telecommunications sector has emerged as a central industry in developed nations, but it faces a significant challenge – customer churn, where valuable customers switch to competitors. With technological advancements and increased competition, telecom companies are exploring three key strategies: acquiring new customers, upselling to current customers, and prolonging customer retention [2]. To do this effectively, they rely on data encompassing customer information over time.

This study primarily centers around the utilization of tree-based and regression-based machine learning techniques and models to forecast customer attrition within the telecommunications sector. The Decision Tree algorithm and Logistic Regression take center stage in crafting an efficient prediction model for customer churn [2].

In today's fiercely competitive market, businesses must constantly adapt and innovate, implementing new strategies and tactics to expand their global customer base for sustained success. This article undertakes an exhaustive exploration of various machine learning algorithms applied in the prediction of customer attrition within the telecommunications industry[3]. It rigorously categorizes and evaluates these research papers, scrutinizing their unique attributes, research methodologies, and specific machine learning approaches employed.

Within the context of Customer Relationship Management (CRM), the system incorporates modules for Cross-Selling, Up-Selling, Customer Retention, and New Customer Acquisition. Cross-Selling strategies aim to target customers who typically do not purchase a particular product and actively promote it to them [4]. On the other hand, Customer Retention strategies primarily revolve around preserving and nurturing the existing customer base within an organization.

The emergence of innovative algorithms has ushered in new opportunities for advancing churn prediction models, offering the potential for further enhancements in model accuracy and the precise identification of churn-prone customers [11] Consequently, these developments have the potential to empower industries in maintaining a minimal churn rate, thereby strategically enhancing their business operations.

2 LITERATURE REVIEW

Abdelrahim Kasem Ahmad employed a variety of machine learning techniques, including XGBOOST, Random Forest, and Decision Tree, to assess their performance. Through this analysis, it became evident that the Area Under the Curve (AUC) served as a crucial metric for evaluating model performance, with higher AUC values indicating superior results. Remarkably, the XGBOOST technique yielded the highest AUC value at an impressive 93%, surpassing all other methods. This outcome highlights XGBOOST as the most successful approach for achieving the best results in Abdulrahim’s study [1].

Abhishek Gaur leveraged a range of machine learning techniques, including Logistic Regression, SVM, Random Forest, and Gradient Boosting, to tackle the challenging task of predicting churn in the future. This comprehensive analysis involved several critical steps, such as feature selection and data division into training and testing sets. The AUC value was found to be 84%, Significantly, the results unveiled that Gradient Boosting emerged as the top-performing technique, exhibiting the highest AUC value. It stood out as the clear winner, delivering the most exceptional results among all the methods considered in Abhishek's study [10].

CV Krishnaveni and AV Krishna Prasad's research delved into diverse churn categories while crafting distinct predictive models. Employing an array of methodologies, including Decision Tree, Random Forest, and XGBOOST, they conducted extensive Exploratory Data Analysis (EDA). This encompassed tasks like addressing unique values and managing missing data.

Their findings unveiled that XGBOOST outperformed the other techniques, providing the highest accuracy. Consequently, they concluded that, in comparison to alternative approaches, XGBOOST emerged as the most suitable model for their churn prediction endeavors [5].

In this research paper, two distinct datasets were meticulously examined. The IBM Watson dataset, comprising 7033 observations and featuring 21 attributes, and the cell2cell dataset, encompassing a substantial 71,047 observations and 57 attributes, were both subjected to visualization using the Orange software [4].

The primary objective of this study was to identify the most accurate predictive model for churn prediction within the telecom domain while also pinpointing the key factors driving customer churn. To achieve this, the researchers implemented three predictive models, namely Naïve Bayes, SVM, and decision tree, using the Matlab platform. The performance of these models was rigorously assessed using the Area Under the Curve (AUC) metric. Remarkably, the AUC values for the IBM dataset were found to be 0.82, 0.87, and 0.78 for Naïve Bayes, SVM, and decision tree, respectively. On the other hand, for the cell2cell dataset, the AUC values stood at an impressive 0.98, 0.99, and 0.98 for the same models [4].

Notably, the AUC achieved through the SVM algorithm surpassed the results of prior studies, signifying its effectiveness in churn prediction [4]. Furthermore, the research revealed that churned customers exhibited common service patterns, suggesting that telecom companies have the potential to identify predictive factors and proactively retain their customer base.

Based on analysis of the AUC values calculated, the most accurate model appears to be gradient boosting, with an AUC score of 84.57%. The ROC curve provides valuable insights into the balance between sensitivity and specificity in our model's performance. Sensitivity measures how effectively the model identifies positive examples, while specificity gauges its ability to correctly classify negative examples . To elaborate on AUC, it represents the area under the ROC curve, and it quantifies the probability that our classifier will rank a randomly selected positive instance higher than a randomly chosen negative one. Therefore, with an AUC score of 84.57%, gradient boosting emerges as the method that excels in distinguishing between positive and negative instances, making it the most accurate model in our evaluation [10].

The findings stemming from our study unveiled that LGBM and XGBoost exhibited the highest levels of accuracy, standing at an impressive 95.74%. Following closely, the RF algorithm achieved the second-highest accuracy rate, registering at 95.38%. Notably, XGBoost demonstrated exceptional performance across a spectrum of evaluation metrics, including precision, recall, specificity, geometric mean (g-mean), ROC score, and Matthews Correlation Coefficient (MCC). Consequently, it is evident that the XGBoost classifier emerged as the optimal choice for our primary objective of predicting churn customers. Furthermore, we observed commendable performance from Gradient Boosting, LGBM, and RF algorithms, reinforcing their viability in our experimental context [12].

The ultimate selection of parameters for the XGBoost model was meticulously crafted, prioritizing the attainment of the highest achievable accuracy, which hovered around 80.5 percent. This process involved a meticulous tuning of each parameter, with an added layer of refinement through the use of grid search to fine-tune these parameters further. Interestingly, a new introduced a novel model into the equation, the resulting accuracy, while marginally superior, still remained in close proximity to that achieved by a logistic regression model [9].

The domain of telecom churn prediction has emerged as a dynamic and evolving field of research, addressing the crucial objective of retaining valuable customers. In recent times, there has been a notable surge in the application of various Machine Learning models to diverse telecom datasets, both in the public and private domains. This article seeks to contribute by offering an extensive survey of the array of machine learning techniques employed within the timeframe spanning from 2000 to 2018.

Furthermore, this paper delves into the landscape of available public and private telecom churn datasets while shedding light on the major challenges faced within the telecom industry. It is noteworthy that a significant upswing in standard research papers occurred in 2017 and 2018, indicating heightened interest and activity in this field during those years.

Presently, hybrid ensembles have gained substantial popularity due to their enhanced predictive capabilities and significant relevance. It provides a comprehensive summary of the various churn prediction endeavors undertaken from 2000 to 2018, encapsulating the evolving landscape of research in this domain [6].

In the competitive telecom industry, preventing customer churn is a top priority for CRM. Researchers are investigating the factors behind churn and how to address them, often using decision tree models. The research developed a churn prediction model, outperforming others, especially when using deep learning like CNN, benefiting telecom companies. In the future, we aim to enhance predictions with advanced methods like sentiment analysis and AI, and study evolving churned customer behavior for trend analysis [8].

3 RESEARCH PAPER STATISTICS

Developing an efficient and effective customer churn prediction model is a fundamental requirement for companies. Various modelling techniques are employed to forecast customer churn across different organizations. Figure 1 illustrates the distribution of research papers and techniques in this field. Below, we present a summary of customer churn prediction models/techniques, with a focus on the most commonly utilized methods: LR, DT, SVM, and RF, each with distinct contributions to churn prediction

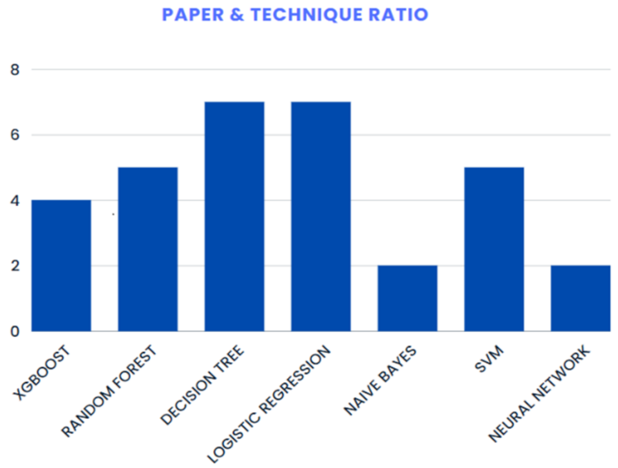

Figure 1. Techniques

In the domain of churn prediction studies within the telecom industry, the choice of datasets plays a pivotal role in shaping the outcomes and insights derived from these endeavors. This pie chart, featured in this review paper, provides an illuminating glimpse into the distribution of datasets used across various research efforts. It underscores the prevalence of telecom company data, Cell2cell, and IBM Watson datasets, each contributing significantly to the body of knowledge in telecom customer churn prediction. The distribution also highlights the diverse sources of data used by researchers, emphasizing the multifaceted nature of this field of study. As we delve deeper into the contents of this review, we will explore how these datasets have been leveraged to develop predictive models and gain insights into customer behavior and attrition trends.

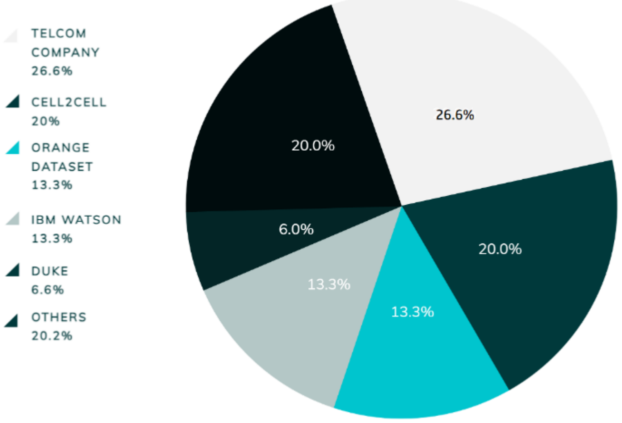

Figure 2. Dataset Distribution in Churn Prediction Studies

4 RESEARCH PAPER ANALYSIS

This table presents a comprehensive overview of various churn prediction models discussed in the referenced papers. It outlines the models, their corresponding reference numbers, and their reported accuracies, providing researchers and practitioners valuable insights into the performance of these models in the context of telecom customer churn prediction.

Models	Reference No.	Accuracy
Logistic Regression	[2]	0.78 precision of weighted average
	[5]	89% of accuracy
	[9]	78.89% of accuracy
	[10]	0.82 AUC value on average model
	[12]	Accuracy is low
Decision Tree	[1]	79.1% using Not-Offered dataset
	[4]	98% of AUC
	[5]	91% of accuracy
	[7]	To be used for churn probability
	[12]	Accuracy is much greater than logistic regression
XGBoost	[1]	89% AUC using XGBoost
	[2]	0.78 precision of weighted average
	[5]	93% of accuracy
	[9]	80.05% of accuracy
Random Forest	[1]	83.4% using Not-Offered dataset
	[2]	0.79 precision of weighted average
	[5]	65% of accuracy
	[10]	0.81 AUC value on average model
SVM	[4]	AUC for SVM (99%)
	[7]	Using low ratio for true churn rate
	[10]	0.79 AUC value an underperforming value
Naïve Bayes	[4]	83% AUC for Naïve Bayes
	[7]	High dimension data is transformed to low dimension
Evolutionary learning data mining technique	[7]	Not to be used on high dimension and large dataset
Gradient Boosted Tree	[10]	0.84 AUC value an best model
GSM (B)	[1]	85.5% using Not-Offered dataset

Table 1.0 Models and their accuracy in the paper

5 RESULT AND CONCLUSION

In this section, we present the results of our analysis and discuss the key findings from the literature review and model techniques of churn prediction in the telecom industry.

5.1 RESULTS

Our research compiled a comprehensive overview of various churn prediction models discussed in the referenced papers. We analysed different models, their respective reference numbers, and reported accuracies. Here are the key results grouped by models:

- Logistic Regression: Logistic Regression was employed in several studies, and its accuracy varied across them. The reported accuracies ranged from 65% to 89%. Notably, one study achieved 78.89% accuracy using Logistic Regression [9].

- Decision Tree: Decision Tree models also demonstrated varying accuracy in different studies. Accuracy rates ranged from 79.1% to 98%. In one study, Decision Tree was found to be much more accurate than Logistic Regression [1][4].

- XGBoost: XGBoost, a popular ensemble learning technique, showed impressive results. Accuracy percentages ranged from 80.05% to 93%. One study achieved an outstanding AUC of 89% using XGBoost [1].

- Random Forest: Random Forest models yielded accuracy rates between 65% to 83.4% across studies. The AUC value reached 81% in one study [10].

- Support Vector Machine (SVM): SVM demonstrated exceptional performance with an AUC of 99% in one study. However, another study reported SVM as an underperforming model with a 0.79 AUC value [4][10].

- Naïve Bayes: Naïve Bayes achieved ssan AUC of 83% in one study, showcasing its potential in churn prediction [4]

- Gradient Boosted Tree: Gradient Boosted Tree stood out with an AUC value of 84.57% in a specific study [10].

- GSM (B): One study reported an accuracy of 85.5% using GSM (B) models [1].

5.2 CONCLUSION

Churn prediction in the telecom sector involves diverse model performance depending on the dataset and research goals. Some studies found XGBoost and SVM to excel, while others favored Logistic Regression and Random Forest for high accuracy. The significance of AUC as a metric emerged, where advanced machine learning methods often resulted in higher AUC values, underscoring their effectiveness. The quality and size of datasets significantly impact model performance, with extensive datasets leading to better results, especially evident in studies with high AUC values. Ensemble learning, like XGBoost and Random Forests, effectively enhances prediction accuracy by combining multiple models' strengths. In summary, churn prediction in telecom is intricate, necessitating careful model selection and data preprocessing. Advanced techniques, particularly ensemble methods such as XGBoost and Gradient Boosted Tree, have shown promise for achieving superior accuracy, with model choice contingent on specific dataset and research objectives.

6 REFERENCES

[1] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. *J Big Data* **6**, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

[2] V. Kavitha , G. Hemanth Kumar , S. V Mohan Kumar , M. Harish, 2020, Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 05 (May 2020)

[3] Rani, K. Sandhya and ., Shaik Thaslima and ., N.G.L. Prasanna and ., R.Vindhya and ., P. Srilakshmi, Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression (JUNE 10, 2021). International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-9, Issue-4, July 2021, https://doi.org/10.21276/ijircst.2021.9.4.6, Available at SSRN: https://ssrn.com/abstract=3902033

[4] Ebrah, K. and Elnasir, S. (2019) Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms. *Journal of Computer and Communications*, **7**, 33-53. doi: 10.4236/jcc.2019.711003.

[5] Krishnan R , CV Krishnaveni and AV Krishna Prasad, Telecom Churn Prediction using Machine Learning, 2022, https://doi.org/10.30574/wjaets.2022.7.2.0130

[6] J. Pamina, T. Dhiliphan Rajkumar, S. Kiruthika, T. Suganya, Femila.F, Exploring Hybrid and Ensemble Models for Customer Churn Prediction in Telecom Sector, 2019,Retrieval Number: A9170058119/19©BEIESP DOI: 10.35940/ijrte.A9170.078219 Journal Website: www.ijrte.org

[7] Nadeem Ahmad Naz, Umar Shoaib and M. Shahzad Sarfraz, A Review on Customer Churn Prediction Data Mining Modeling Techniques, 2018, IJST. 10.17485/ijst/2018/v11i27/121478

[8] J.K.Kiruthika , B.S.Neshamoney , L.Madhan Kumar , V.Mugunthan, Churn Prediction in Telecom sector using Deep Neural Network with Flask Application, 2021,IOP, **DOI** 10.1088/1757-899X/1166/1/012060

[9] Miss.Priyanka Parmar, Telecom Churn Prediction Model using XgBoost Classifier and Logistic Regression Algorithm, 2021,IRJET.

[10] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), Bhopal, India, 2018, pp. 1-5, doi: 10.1109/ICACAT.2018.8933783.

[11] Y. Bharambe, P. Deshmukh, P. Karanjawane, D. Chaudhari and N. M. Ranjan, ""Churn Prediction in Telecommunication Industry"," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-5, doi: 10.1109/ICONAT57137.2023.10080425.

[12] Taskin, N. (2023). CUSTOMER CHURN PREDICTION MODEL IN TELECOMMUNICATION SECTOR USING MACHINELEARNING TECHNIQUE (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-506134.