

Telecom Industry Customer Attrition Prediction with Machine Learning Approach

Deepika Bansal¹, Bhoomi Gupta² and Sachin Gupta³

^{1,2} Department of Information Technology

³ Department of Computer Science Engineering

Maharaja Agrasen Institute of Technology

New Delhi 110086, India

Abstract. In order to keep current customers, the organization in the telecommunications sector must focus on one of the very important research areas: customer churn identification. Customer attrition known as churn occurs when rivals discontinue selling certain products or services, or possibly when network issues arise. Customers often tend to abandon their service subscriptions in these kinds of circumstances. The churn rate significantly influences the client's lifetime value since it influences both the company's potential income and the length of service. The corporations are searching for a model that can forecast client attrition as the revenues are directly impacted. Machine learning techniques are used in the proposed model. For a more sustainable business environment, companies need to understand and effectively manage customer attrition. Reducing network issues, resource allocation optimization and waste reduction can improve customer satisfaction and thus decreasing churn rates. Optimizing service provisioning and resource allocation helps in promoting sustainable decision making and enhances revenue generation by development of predictive models utilizing machine learning concepts and methods for artificial intelligence. It empowers organization to build long lasting customer relationships and minimizing environmental impact of their operations.

Keywords: Customer Churn, Telecom, Predictive Analysis, Optimization, Client Retention

1 Introduction

In industrialized nations, the telecommunications industry has emerged as one of the most important economic sectors. These service industries struggle, particularly as a result of client loss to competitors or consumer attrition. Scientific development and a surge in operators led to more opposition [1]. Businesses are putting in a lot of effort and depending on complex strategies to survive in this cutthroat market. The attrition of customers has a profound impact on the Indian telecom service providers for several reasons. There is an intense competition among service providers creating a highly competitive environment where customers have a plethora of options to choose from. Price sensitivity among Indian consumers is a major factor, with even slight differences in pricing leading service porting by clients in search of better terms. The introduction of Mobile Number Portability (MNP) provides an instant opportunity to change service providers while keeping the same phone numbers, making it easier for them to consider the switch. Additionally, the lack of significant differentiation between telecom operators in terms of core services like calling and data further contributes to customer churn, as customers may perceive little added value beyond price. Some telecom companies prioritize acquiring new customers over retaining existing ones, leading to a cycle of churn as efforts are focused more on attracting new subscribers rather than ensuring customer satisfaction and loyalty. This approach can result in a constant need to replace departing customers with new ones. Furthermore, evolving customer needs and expectations, driven by the emergence of data-driven services and Over-The-Top (OTT) platforms like WhatsApp and Netflix, require telecom operators to adapt their offerings to meet changing demands. Failure to keep pace with these evolving trends can lead to a loss of customers seeking more flexible and feature-rich services from alternative providers.

Customer turn over becomes a big problem and causes a considerable loss of telecom services to survive in this market [2]. Customer turn over becomes an immense problem leading to a considerable loss of telecom services. Three key techniques have been implemented to increase profitability: add untapped customers, retain the present clients, and eliminate churn. During the investigation by comparing the respective impact of these factors on the Return on Investment (RoI) values, the experiments have shown that the last method yields maximum benefits[3][4]. The remainder of this paper is organized as follows: Section 2 describes the literature survey, Section 3 explains methodology, results are discussed in Section 4 and Conclusion in Section 5.

2 Literature Survey

Several studies have shown that the Customer attrition as seen in the Indian telecom industry results from various factors as highlighted by researchers. These factors include:

1. **Intense Competition and Price Wars:** The telecom sector in India is a highly competitive market, leading to reflective pricing models and innovation in the strategies for marketing, which can prompt customers to switch providers.[14]
2. **Shift in Customer Usage Patterns:** It has been observed that a simple pattern showing a significant change in service usage happens before churn occurs, a term popularly known as revenue churn, which in turn potentially leads to customer attrition from the service provider..
3. **Service Quality and Offers:** Better subscription offers, service quality, and customer experience play a significant role in customer churn. Factors like digital experience, service delivery, consumer interaction, monetary benefits, and product promotions are crucial in influencing churn rates.[15]
4. **Tariff Plan Changes and Product Offerings:** Changes in tariff plans, product offerings, and service types can also impact customer churn. For instance, dissatisfaction with services or better offers from competitors can lead to churn.
5. **Customer Expectations and Loyalty:** Customers today are more demanding and comparison shoppers, expecting better services at competitive prices. This increased demand reduces customer loyalty and makes them more likely to switch providers.

Numerous techniques, like as machine learning, data mining, and hybrid technologies, have been employed to predict client attrition. Mentioned methods help businesses recognise, foresee, and hold on to churning clients. The methods can also be helpful for the customer relationship management, and strategic decision making for the industry. A majority of these methods made use of machine learning decision tree algorithms, which are a recognized method for figuring out customer turnover but not ideal with difficult situations [5]. However, the study demonstrates that a reduction in data quantity increases model accuracy. Some other techniques like data mining are also employed for consumer forecasting and historical research. In addition to other frequent prediction techniques like neural networks, rule-based learning, and decision trees, regression tree techniques were discussed [6].

In this paper, we employ several techniques, including Random Forest, and Logistic Regression augmented with XGBoost, to obtain precise values and forecast customer attrition [7][8]. Here, we model using a dataset that has undergone training and testing, giving the highest number of accurate results. The suggested model for churn prediction is depicted and its phases are described in Fig. 1. Data preparation is done in the first phase when we filter data and transform it into a comparable shape before choosing features [9].

In the following stage, methods including Random Forest, XGBoost, and Logistic Regression (LR) are used for forecasting and categorizing data. With domain specific collected data, we apply training and test the model while observing and analyzing client behavior. Customer attrition prediction is performed using machine learning in several other recent works as well [10][11][12][13].

3 Methodology

3.1 Dataset

The data set serves as the foundation for everything and must have sufficient data to enable machine learning of the issue. Datasets can be constructed or created from discarded data that is accessible online. Building a dataset that makes sense and offers direction on how to react depending on real-time inputs for the issue is one of the issues we face.

Some datasets suitable for telecom churn prediction in India and other regions include:

1. Telecom Churn Prediction Dataset on GitHub: This dataset offers a vanilla classification model for Telecom Churn predictions Asian Market including India [16]
2. Telecom Customer Churn Dataset on Kaggle: For analysis of customer-level data made available by a telecom leader to build predictive models identifying customers at high risk of attrition [17]
3. Telecom Churn Indian and South East Asian Market Dataset on Kaggle: Similar to the previous dataset, this project involves analyzing customer-level data to predict high-risk churn customers in the Indian and South East Asian markets [18]

The features included in telecom churn prediction datasets typically encompass a range of factors that influence customer behavior and the likelihood of churn. Some common features found in these datasets are:

- Information about customers' demographics such as age, gender, location, and income level.
- Service Usage Details: Data related to how customers use telecom services, including call duration, data usage, messaging patterns, and service subscriptions.
- Payment History: Details about customers' payment behavior, billing history, payment methods, and transaction records.
- Social Network Analysis (SNA): Utilizing customer social network data to understand relationships and influences that may impact churn prediction.
- Customer Interaction: Records of customer interactions with the telecom company, including complaints, inquiries, and feedback.
- Tariff Plan Information: Details about the tariff plans subscribed to by customers and any changes in pricing or offerings.
- Customer Satisfaction Metrics: Feedback scores, surveys, or sentiment analysis results reflecting customer satisfaction levels.
- Network Coverage: Information on network coverage quality and availability in different regions.

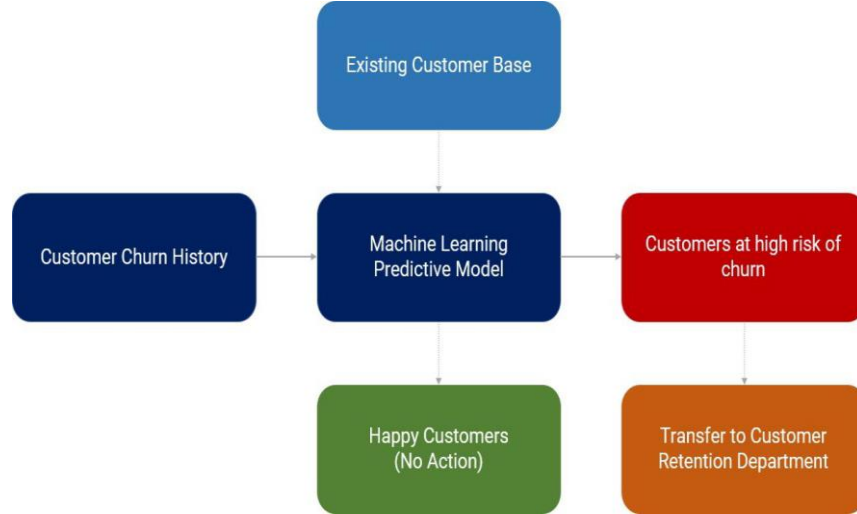


Fig. 1. Proposed Flow Diagram for customers attrition prediction

3.2 Data Preprocessing

A data set has N rows and contains a collection of features. Different forms exist for many values. In a dataset, there could be redundant or null values that cause some loss in accuracy and dependencies.

A single variable, such as gender, is noted using a variety of formats since the data were gathered from several sources. For example, M/F or Male/Female is used to indicate gender. A three-dimensional image should be converted to a two-dimensional format in order to prevent data that is noisy, has null values, and incorrect sizes as data shown since the machine can only interpret numbers 0 and 1. Panda's tabular data and OpenCV for pictures are both capable of doing data cleaning, data filtering and noise removal.

3.3 Feature Selection

Feature selection process of machine learning helps to select the relevant components from the collected data based on domain knowledge. The study employed a dataset with numerous characteristics, from which we identified those deemed most crucial for enhancing performance assessment and facilitating decision-making. Classification performance increases if the set of data only contains useful and highly predictable variables. As a result, classification performance is improved when the irrelevant characteristics number is kept to a minimum and only important features are used.

3.4 Algorithms used

A. Random Forest

We use Random Forest to forecast if a user is likely to unsubscribe his membership. Random Forest algorithms use decision trees to forecast if a customer would terminate their subscription. Each random forest decision comprises several separate decision trees of varying depths. A decision tree pinpoints a certain group. The classifier used for a certain customer will be the one with maximum votes. Decision trees' behaviour may vary depending on the dataset they are being trained on. For this reason, bagging is used. One method is to train the decision trees using a random sample from the dataset. The user selects the k nearest neighbours, indicated by the letter K in the classifier's name, and they represent.

B. `naive_bayes.MultinomialNB`

The simplest yet effective naive Bayes classifier works excellently in multinomial models. During the classification of data with discrete features, including data based on word counts towards text classification, the multinomial Naive Bayes classifier is expected to work well. Integer feature counts are typically utilized for multinomial distribution but some other features including fractional counts like tf-idf have also been found to be effective. Accordingly, using the multinomial Naive Bayes classifier makes sense.

C. `SGDClassifier`

For logistic regression and other such classifiers which are linear in nature, a model based on SGD training is also used. Essentially, the estimator in this technique uses regularised models based on stochastic gradient descent (SGD) learning, upgrading the model with a schedule of diminishing strength as we progress (termed **learning rate**). This strategy following a partial fit in SGD method allows the learning of online or out-of-core mini batch. It is essential to note that the default learning rate schedule can be applied to data having characteristics of zero unit variance and mean.

D. `KNeighborsClassifier`

This classifier uses, as the name implies, learning based on the k nearest neighbours. The appropriate value of k is determined by the data. The study demonstrates that decreasing the data enhances the decision tree's accuracy. Consumer prediction and historical analysis are two uses for data mining techniques. Alongside various famous data mining approaches such as decision trees, rule-based learning, and neural networks, regression tree techniques were discussed.

4 Results

To improve the accuracy further data balancing was performed because the data was imbalanced. Now label encoding and one hot encoding is used and in addition, SMOTE algorithm is used for managing the imbalance. Many firms, especially the telecom service providers, will benefit from the relevance of this study for attrition prediction in obtaining robust revenue and profitable earnings. Due to the difficulty in estimating customer churn in the telecom sector, companies aim on prioritizing on maintaining their existing clients above attracting the new ones.

Various performance measures like Accuracy, Recall, F1-score and Precision have been calculated and shown in Table 1.

Table 1. Various performance measures evaluated.

Classifier	Accuracy	Recall	F1-Score	Precision
Random Forest	0.80	0.97	0.87	0.80
Naïve_Bayes	0.89	0.91	0.90	0.89
SGD Classifier	0.76	0.80	0.80	0.85
K-Neighbors	0.83	0.86	0.90	0.85
Voting Classifier	0.90	0.90	0.91	0.92
SMOTE	0.97	0.95	0.96	0.96

Key Observations:

1. Accuracy: SMOTE achieves the highest accuracy (0.97), indicating it correctly predicts churn in 97% of the cases. The next best model, Voting Classifier, has an accuracy of 0.90.
2. Recall: While both SMOTE and Random Forest have the highest recall (0.97), signifying they identify most churning customers accurately, it's a tie.
3. F1-Score: SMOTE has the best F1-Score (0.96), which balances precision and recall, demonstrating its effectiveness in churn prediction.
4. Precision: SMOTE has the highest precision (0.96), meaning most of the customers it identifies as churning actually churn.

SMOTE stands out as the most effective technique for telecom service attrition prediction in this scenario. It achieves superior performance in accuracy, F1-Score, and precision. While Random Forest ties with SMOTE in recall, SMOTE's well-rounded performance across all metrics makes it the preferable choice.

5 Conclusion and Future Work

The applicability and variety of three tree-based algorithms in this kind of application led to their selection. We will get more accuracy by combining Logistic regression with Random Forest. By using the SMOTE Algorithm we get good results. And accuracy is above 90 per cent. The selection of three tree-based algorithms was on the basis of their suitability and variety for this kind of application. We will obtain greater accuracy by utilizing Logistic Regression and Random Forest. Using the SMOTE algorithm, we achieve decent performance.

Massive amounts of data are produced in today's environment, and businesses rely on this data to advance. To be useful to companies, massive amounts of unstructured data stored in the cloud need to be prepared and analyzed. In the last ten years, data-driven technology has transformed both our personal lives and our professional lives. Drilling the data is a goal of data science in order to find hidden possibilities and provide value. At all levels of business, it has played a crucial role.

Potential research avenues in the area of customer attrition prediction, beyond the present study, could include:

- Causal Analysis of Churn Factors: Conducting in-depth studies to understand the causal relationships between various factors and customer churn. This research can help uncover the root causes of churn and guide targeted interventions.
- Segmentation Analysis: Exploring different customer segments based on behavior, preferences, and churn patterns to create more effective targeted retention strategies. It will also help to realize the individual needs of varied customer groups leading to tailored retention approaches.
- Longitudinal Studies: Undertaking longitudinal studies to track customer behavior over time and identify patterns that precede churn. This research can provide critical information into the ever changing dynamics of customer relationships with telecom service providers.
- Qualitative Research: Incorporating qualitative research methods like interviews, focus groups, or surveys to delve deeper into customers' perceptions, emotions, and experiences related to churn. Qualitative insights can complement quantitative data for a comprehensive understanding.
- Predictive Analytics Enhancements: Advancing predictive analytics techniques by integrating real-time data analysis for immediate action on potential churn signals. Research focusing on enhancing the accuracy and timeliness of churn predictions can lead to more proactive retention strategies.
- Ethical Implications and Fairness: Investigating the ethical implications of using sensitive customer data in churn prediction models to ensure fairness,

transparency, and responsible practices. Research in this area can address concerns related to biases and discrimination in machine learning algorithms.

References

1. Olle Olle, D. and Cai, S.: A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of e-Education, eBusiness, e-Management and e-Learning*, 4(1), (2014).
2. Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A.: Telecommunication subscribers' churn prediction model using machine learning. In *Eighth international conference on digital information management*, 131-136 (2013).
3. Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M.: A proposed churn prediction model. *International Journal of Engineering Research and Applications*, 2(4) 693-697 (2012).
4. Rani, K. S., Thaslima, S., Prasanna, N. G. L., Vindhya, R., & Srilakshmi, P. : Analysis of customer churn prediction in telecom industry using logistic regression. *International Journal of Innovative Research in Computer Science & Technology (IJRCST) ISSN, 2347-5552* (2021).
5. Balasubramanian, M., & Selvarani, M.: Churn prediction in mobile telecom system using data mining techniques. *International Journal of scientific and research publications*, 4(4), 1-5 (2014).
6. Gaur, A., & Dubey, R.: Predicting customer churn prediction in telecom sector using various machine learning techniques. In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, 1-5 (2018).
7. Chouiekh, A.: Machine learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco. In *2017 Intelligent Systems and Computer Vision (ISCV)*, 1-8 (2017).
8. Balasubramanian, M., & Selvarani, M.: Churn prediction in mobile telecom system using data mining techniques. *International Journal of scientific and research publications*, 4(4), 1-5 (2014).
9. Kavitha, V., Kumar, G. H., Kumar, S. M., & Harish, M.: Churn prediction of customer in telecom industry using machine learning algorithms. *International Journal of Engineering Research & Technology (IJERT)*, 9(5), 181-184 (2020).
10. Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H.: Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, 100342, (2024).
11. Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E.: Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management*, 7(1), 7-16 (2024).
12. Akhmetshin, E., Fayzullaev, N., Klochko, E., Shakhov, D., & Lobanova, V.: Intelligent Data Analytics using Hybrid Gradient Optimization Algorithm with Machine Learning Model for Customer Churn Prediction. *Fusion: Practice and Applications*, 14(2), 159-59 (2024).
13. Saha, S., Saha, C., Haque, M. M., Alam, M. G. R., & Talukder, A.: ChurnNet: Deep Learning Enhanced Customer Churn Prediction in Telecommunication Industry. *IEEE Access*, (2024).
14. Geetha, M., Abitha Kumari, J.: Analysis of churn behavior of consumers in Indian telecom sector. *Journal of Indian Business Research*. 4, 24–35 (2012). <https://doi.org/10.1108/17554191211206780>.
15. Gowd, S., Mohite, A., Chakravarty, D., Nalbalwar, S.: Customer Churn Analysis and Prediction in Telecommunication Sector Implementing Different Machine Learning Techniques. In: Manza, R., Gawali, B., Yannawar, P., and Juwono, F. (eds.) *Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)*. pp. 686–700. Atlantis Press International BV, Dordrecht (2023). https://doi.org/10.2991/978-94-6463-196-8_52.

16. "Debmalya92/Telecom-Churn-prediction: Two different approach to predict churn customers and finding out important variables that drives churn," GitHub, 2019. [Online]. Available: <https://github.com/debmalya92/telecom-churn-prediction>. [Accessed: 03-Mar-2024]
17. "Telecom customer churn dataset," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/shivam131019/telecom-churn-dataset>. [Accessed: 03-Mar-2024]
18. Ahmad, A.K., Jafar, A., Aljoumaa, K.: Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*. 6, 28 (2019). <https://doi.org/10.1186/s40537-019-0191-6>.