# AN OVERVIEW OF CUSTOMER CHURN PREDICTION IN TELECOM INDUSTRY

*A Literature Survey*

[1]Afroz Chakure,[2]Amarkumar Belkhede,[3]Saksham Shrimali,[4]Siddharth Garud,[5]Seema Patil
[1]Student ,[2]Student ,[3]Student, [4]Student, [5]Assistant Professor

[1]School of Computer Engineering and Technology,
[1]MIT World Peace University, Pune, India.

***Abstract :*** In Telecom Industry customer churn is a big issue and one that impacts their revenue. When customers start to leave a service or subscription, it increases the expenditure for these companies. Businesses have found that acquiring new customers costs them nearly six times more money than retaining existing ones. Therefore, preventing customer churn becomes important when companies are trying to grow their business. The analysis of Customer Behaviour using Machine Learning techniques does provide an effective solution to the problem by predicting which customers are more likely to leave the service or subscription. Predictive analysis of customer behaviour not only helps companies fix issues with their service but also helps them add new features and products so as to keep the customer engaged. The present work provides an overview of the latest works in the field of Customer Churn prediction. Our aim is to provide a simple path to make the future development of novel Churn prediction approaches easier.

***IndexTerms* - Churn Analysis, Telecom Churn, Customer Churn Prediction, Customer Retention, Classification, Machine Learning.**

## I. INTRODUCTION

In modern day business, it is important for companies to stay competitive, be innovative and generate profits for their investors. But being competitive and innovative alone doesn't guarantee healthy returns for businesses. The Telecom Industry is a highly saturated business and is well known to adopt strategies [14] like (1) acquiring new customers, (2) upselling the existing customers and (3) increasing the retention period of customers, to generate revenue. But Nowadays with Multiple service providers offering the same service at highly competitive rates, means that customers have a lot more incentive to switch from one provider to another.

To stay relevant it is imperative for companies in the Telecom sector to understand the behaviour of customers who are using their service and find out which all customers are more likely to leave the service. Predictive modeling tends to provide a solution to the problem by using advanced Machine Learning Algorithms that help analyse and predict customers sentiments. It helps companies to pivot and fix issues with their service faster by understanding customer behaviour. Companies can make use of existing data about their customers to find insights about their interests, their gender and region distribution, along with an understanding about which services are popular and which are not.

Implementation of a prediction model helps telecommunication companies to discover trends and future behaviours that allows them to take smarter decisions based on knowledge extracted from data [15]. Customer retention has a high Return on Investment (ROI) for businesses as acquiring a new customer costs companies six times higher than the cost of retaining the customer which is likely to churn [2].

Customer churn tends to define the business strategy for many companies as it not only affects the profits of such companies but also affects their brand value [4]. The telecom industry has two models of payment for their services namely prepaid and postpaid. While customers in the postpaid service cannot leave the provider anytime, it is not the same with prepaid users. Prepaid users can churn anytime without any prior notice to the provider. Depending on the market region these telecom companies operate in, it would be beneficial if they understand their customers requirements and the services they prefer beforehand to avoid major losses in the future.

In this paper we try to break down research in the field of Customer Churn Prediction in Telecom Sector by exploring various literatures that are available on this topic. The findings from each of the papers has been provided and a summary of results with respective algorithms used is given.

## II. LITERATURE SURVEY

There are many methods to solve the problem of Customer Churn prediction in the Telecom Sector. Most of the research in this field is focused on applying multiple Machine Learning algorithms to make predictions and later comparing the results. Few papers have gone so far to devise newer algorithms to apply to this problem while some make use of existing algorithms with better tweaking of parameters.

The paper by NGURAH P. et al. [1] tries to solve the problem of customer churn by using Deep Neural Network (DNN), Random Forest and Extreme Gradient Boosting (XGBoost) algorithms on IBM customer dataset having 7040 rows and 21 fields. The architecture of the DNN model is provided along with comparison with previously studied models like Random Forest and Extreme Gradient Boosting (XGboost). The study finds that the proposed DNN model gave an accuracy of 80.62% with a processing time of 64 seconds while algorithms like Random Forest and XGBoost gave accuracy of 77.87% and 76.45% with processing times of 529 seconds and 175 seconds respectively.

In the research conducted by Ahmad et al. [2] on Customer Churn prediction by building and testing models on SyriaTel Telecom company dataset using Spark environment. The paper describes the steps in feature engineering, feature transformation, and selective approach to make features ready for the Machine Learning algorithms. Four tree based algorithms are used namely Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM) and XGBoost algorithm. Out of these 4, XGBoost model gave the best results with AUC value of 93.301% while GBM algorithm came second, and Random Forest and Decision Tree came third and fourth w.r.t AUC values. The authors of the paper found that using Social Network Analysis features enhanced results in predicting churn in telecom.

Gaur A. et al. [3] have focused their research on applying multiple Classification algorithms like Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosted tree and have compared the performance of these models. A brief review of previous literature has been done by the authors and steps involved in building the model are provided along with flowcharts for Churn Prediction Framework and Analysis Steps. Their research found that Gradient Boosting performed best among the four models and results given by Logistic Regression and Random Forest were average while SVM performed the worst in their test. They have used the metric of AUC (Area under the ROC Curve) to determine the correctness of their model.

Paper by Ullah et al. [4] investigated churn prediction in the present market of the telecom industry. CRM is a significant factor for churn prediction to retain the customer and provide offers or services to maintain that group of customers. The dataset was preprocessed using noise removal and feature selection. In [4] churn prediction performed better by using the machine-learning algorithm such as Random Forest, which gives the accuracy of 0.88. Clustering techniques like K-Means have been used by identifying the main factors from the dataset, which are used in predicting the churn customers. Metrics, such as accuracy, precision, recall, f-measure, and ‘receiving operating characteristics’ (ROC) area were used. The results revealed that their proposed churn prediction model produced better churn classification by using the RF algorithm and customer profiling by K-means clustering.

A unique approach was followed by Idris et al. [5] where the authors propose using Genetic Programming (GP) based approach to model the problem of customer churn prediction. Ensemble approach has been considered to give better performance than individual classifiers. Theirs being an imbalanced dataset with fewer instances of the minority class, Random Forest ensemble would prove problematic. On the other hand, flexibility and distinctive features offered by GP, makes it more suitable for classification and in turn, churn prediction. Whereas, Adaboost is a boosting technique that works to combine multiple weak classifiers in order to create a strong one. Dataset used here is the “Orange Telecom” dataset and “Cell2Cell” dataset. Area under the curve (AUC), sensitivity and specificity are the measures used here to evaluate predictor performance. Highest churn prediction accuracy of 0.89 AUC is reported on Cell2Cell dataset.

Hybrid approach is followed by Joolfoo et al. [6] where ANN (Artificial Neural Networks) and KNN (K-Nearest Neighbors) have been used for predicting telecom churn. Previous studies have been compared and a systematic review of different models used previously has been done. This was conducted on a total of 15 papers dating from 2014-2020. They have tried to devise a novel and hybrid approach using the KNN machine Learning algorithm along with Artificial Neural Networks. Accuracy is the proposed measure of performance for this model.

Toderean G. [7] we studied that they have used a dataset consisting of 3333 customer call details which has 21 features and that they have implemented advanced data mining methodology for churn prediction, it has yes/no as depending parameter. Customer dataset has details such as incoming, outgoing calling and voicemail. We studied that the author has implemented PCA (principal component analysis) along with Bayes Network, Support vector machine and Neural Network. For calculating the performance of the algorithm, the author had used area under curve, Bayes Networks 99.10%, Neural networks 99.55% and support vector machine 99.70%.

The problem of customer churn prediction in Big data platforms has been studied by Huang [8]. The goal of the researchers was to prove that big data greatly enhances the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China’s largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC. At a point in time, there were 5 millions of active customers, and the system could generate a draft of prepaid customers who are leaving the service (churn) in next month, which has 0.96 precision for the customer who are at the top 50000 in the prepaid customer list.

The paper on “Behavioral Modeling for Churn Prediction” by Khan M. et al. [9] we studied the beginning indication of churn and developed a churn score, so the company can distinguish the customer who is about to end the service. The authors advanced towards uses brute force to feature engineer which can lead to large numbers of overlapping features from the customers data such as calls, logs, etc. then uses two related techniques to identify the features and metrics that are most predictive of customer churn.To predict subscriber churn, the features are given into series of supervised learning algorithms. A South Asian mobile phone operator consisting of terabytes of data was used for testing the approach when the authors classify the subscribers who were inactive on more than 76 percent of days during the training period as churners, their prediction is correct in 83.9% percent of cases. The good performance of this simple linear discriminant is due to the fact that they have an unbalanced sample, in the sense that 76.6 % of their sample does not churn, and a very simple model that just predicted the majority class (i.e. “not churn”) for all subscribers would achieve 76.6% accuracy. Depending on the algorithm used to predict churn, they achieved accuracy rates of roughly 88.5-89.5%

While using a single algorithm to predict Churn can give good results, It is often found that using multiple or a combination of these algorithms is often the key to producing great results. Hu X. et al. [10] in their paper used a decision tree, a neural network and a combination of both to predict the churn. The data came from customer information of a supermarket from June, 2018 to April, 2019 and a total of 2681 entries were used after preprocessing. The prediction accuracy of the decision tree model is 93.47% and of the neural network model is 96.42%. After removing 21 customers whose churn probability is between 0.4 to 0.5 the accuracy of the combined model is 98.87%. It mainly aimed at the problem that it is difficult for a single model to achieve high accuracy.

Most datasets in Telecom Churn are pretty common and researchers have oftentimes tried to compare their results based on accuracy but accuracy alone is never a good measure of the model. Malviya K. et al. [11] in his paper compared two machine learning algorithms, SVM and Random Forest based on different aspects like accuracy, specificity and sensitivity. Dataset was divided in 75%-25% ratio for training and testing purposes respectively. Only “streaming movies” and “streaming tv” are highly correlated. AUC of Random Forest is 81.24% and of SVM is 79.75%. According to this paper Random Forest performed better in terms of accuracy and specificity, but SVM performed better in terms of sensitivity. The data used 7043 observations and 21 variables extracted from a data warehouse.

Decision Tree Techniques like CART and Random Forest were studied by Shrikhande P. et al. [12] and used for predicting customer churn through which they build classification models and also compared the performance of these models with logistic regression model. The authors followed steps as Data Mining like Data Preparation, Data, Preprocessing, Data Extraction and Decision in that order. The dataset contains 3,333 observations and 21 variables extracted from the data warehouse. When comparing the results it is clear that the predicting churn is more accurate in random forest because it gives 98.31% sensitivity. The true positive rate known as the sensitivity or probability of detection measures the proportion of positives that are correctly identified as churn. Based on the result it can be said that random forest performs better in customer churn prediction because it has better sensitivity as compared to CART and Logistic Regression models

Deep learning techniques like the CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) have been used by S. Kavitha et al. [13] in this paper to predict the churn rate. According to this paper customer churn prediction performs a crucial role in the telecommunication industry due to the smartphone dominated era. In previous studies, machine learning techniques used, faced the problem due to imbalanced data. In this paper, deep learning technique removed the missing values and redundant data to rectify it. Accuracy of 80% was shown by CNN algorithm and 74% by RNN algorithm. It clearly shows that CNN gives better results in detecting customers who have a tendency to shift to the competitors. It can be helpful to retain customers and their subscription with the same company.

## III. Results and Conclusion

In this paper we have provided a systematic overview for the problem of customer churn analysis and prediction in the telecom industry using Machine Learning Methods. Churn Analysis helps Telecom companies retain existing customers on the basis of predictive results. A review of past work in the field of Customer Churn prediction in the Telecom Sector was presented and respective advantages and disadvantages were highlighted. Our studies found that the results vary greatly based on different datasets used in research. Also Machine Learning Algorithms play a role in deciding the accuracy of the models. Our aim is to give a clear pathway for future development of Churn Prediction using Machine Learning methods.

## IV. Acknowledgment

We would like to thank MIT World Peace University, Pune and School of Computer Engineering & Technology for giving us the opportunity to research on this topic and work on it as part of our Final Year BE project.

## REFERENCES

**[1]** N. Putu Oka H and A. Setyo Arifin, "Telecommunication Service Subscriber Churn Likelihood Prediction Analysis Using Diverse Machine Learning Model," 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT), Medan, Indonesia, 2020, pp. 24-29, doi: 10.1109/MECnIT48290.2020.9166584.

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

**[3]** 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.

**[4]** I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

**[5]** A. Idris, A. Khan and Y. S. Lee, "Genetic Programming and Adaboosting based churn prediction for Telecom," 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Korea (South), 2012, pp. 1328-1332, doi: 10.1109/ICSMC.2012.6377917.

**[6]** M. B. A. Joolfoo, R. A. Jugumauth and K. M. B. A. Joolfoo, "A Systematic Review of Algorithms applied for Telecom Churn Prediction," 2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), Balaclava, Mauritius, 2020, pp. 136-140, doi: 10.1109/ELECOM49001.2020.9296999.

**[7]** ABrandusoiu, Ionut & Toderean, G. & Beleiu, Horia. (2016). Methods for Churn Prediction in the Pre-paid Mobile Telecommunications Industry. 97-100. 10.1109/ICComm.2016.7528311.

**[8]** Khan, M., Manoj, J., Singh, A., & Blumenstock, J. (2015). Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty*2015 IEEE International Congress on Big Data*.

**[9]** R. Mohanty and K. J. Rani, "Application of Computational Intelligence to Predict Churn and Non-Churn of Customers in Indian Telecommunication," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 2015, pp. 598-603, doi: 10.1109/CICN.2015.123.

**[10]** X. Hu, Y. Yang, L. Chen and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 129-132, doi: 10.1109/ICCCBDA49378.2020.9095611.

**[11]** K. Malviya and S. R. Ali, “Predicting Customer Churn Prediction in Telecom Sector Using SVM AND Random Forest”, 2019 JETIR June 2019, Volume 6, Issue 6. JETIR1906B29.

**[12]** P. A. Shrikhande and Prof. A. Verma, “Performance Enhancement of Customer Churn Prediction in Telecom Sector using Decision Tree Techniques”, 2018 JETIR November 2018, Volume 5, Issue 11. JETIR1811934

**[13]** S.Kavitha, R.Seetha and S.Sathyavathi, “Prediction of Customer Churn in Telecom Industry Using Deep Learning Techniques”, 2020 JETIR 2020 June 2020, Volume 7, Issue 6.

**[14]** Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst Appl. 2002;23(2):103–12

**[15]** I. Brându oiu, G. Toderean, and H. Beleiu, “Methods for churn prediction in the 3re-3aid Mobile telecommunications □ .Industry,” pp. 97–100, 2016