# Telecom Churn Prediction Model using XgBoost Classifier and Logistic Regression Algorithm

**Miss.Priyanka Parmar[1]**

*P.G. Scholar, MASTRE OF CE KITRC, Ahmedabad, Gujarat, India*

**Mrs. Shilpa Serasiya[2]**

*H.O.D, CE Department. KITRC, Ahmedabad, Gujarat, India*

---

***Abstract*:-** We will increase gadget getting to know version which clients who've now no longer completed any usage, both incoming or outgoing - in terms of calls, net etc. over a length of time. A ability purchaser has stopped the use of the offerings for a while; it is able to be too overdue to take any corrective moves to keep them. For e.g., in case you outline churn primarily based totally on a 'two-months 0 usage' length, predicting churn may be vain when you consider that via way of means of that point the purchaser could have already switched to every other operator. For that set of rules we will use dataset carries purchaser-degree records for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, eight and 9, respectively. We can be going to use logistic set of rules to discover climate purchaser is churn or now no longer. To discover a purpose why it's miles churn or now no longer we are able to follow XGBOOST Classifier set of rules to discover which function has affected to grow to be a churn or now no longer. We will first follow facts pre-processing and cleaning the facts the use of pandas and numpy library and carry out EDA (Exploratory facts analysis) Task. After, efficaciously completed EDA Analysis we are able to put into effect Logistic set of rules for prediction version and get higher overall performance and function significance we are able to put into effect XGBOOST Classifier which offer us capabilities which affected to purchaser grow to be a churn or now no longer. For in addition enhancement we are able to do Customer Behavioral Analysis the use of supervised strategies and acquire green facts on big scale.

***Keywords: Telecom Churn, EDA (Exploratory Data Analysis Xgboost (Extreme Gradient Boosting) Classification Algorithms*.**

## 1. INTRODUCTION

Simple terms, customer churn occurs when the consumer wants to completely stop your services and switch to another provider.

Customer churn has a major impact on businesses that rely primarily on subscription revenue. The amount of persons who leave a group over time is referred to it as the churn rate. Today's telecommunications firms are up against a lot of stress. competition today, as every corporation comes with a brand new scheme each month to draw clients. Losing clients in this type of enterprise may be very costly, as it's miles extraordinarily tough to draw NEW clients. Bad carrier opinions on the internet, phrase of mouth opinions are elements that discourage ability clients from becoming a member of a brand new network. It turns into extraordinarily crucial for such corporations, to hold its cutting-edge clientele. Finding out what may be viable elements that reason a purchaser to churn could assist corporations provide higher services custom designed to a specific consumer's needs. At instances it is able to manifest that a precise consumer has a better want for records and the cutting-edge issuer turns out to be too steeply-priced and the consumer switches to any other corporation. Predicting which customers should viable transfer can assist telecom corporations devise a brand new plan for the ones unique clients so one can save you them from leaving the network. In reality this isn't always new, the instant you install a request to stop services, and also you get a name from them with a few new plan to trap you. However, from a non-public experience, as soon as a purchaser is bored stiff together along with your services, and has already made the selection to transfer, it's miles extraordinarily tough to persuade them to stay. It is higher to provide such plans BEFORE they make the selection.

Supervised Machine Learning is not anything however gaining knowledge of a characteristic that maps an enter to an output primarily based totally on instance enter-output pairs. A supervised gadget gaining knowledge of set of rules analyzes the schooling statistics and produces an inferred characteristic, which may be used for mapping new examples. Given that we've got statistics on contemporary and previous purchaser transactions within side the telecom dataset, that is a standardized supervised category hassle that attempts to are expecting a binary outcome (Y/N).

### A. *Existing System*

Customer churn prediction has been performed using numerous strategies, including information mining, device learning, and hybrid technologies. These strategies permit and support companies in identifying, predicting, and retaining churn customers. They also assist industries in CRM and selection making. Most of them used selection bushes in not unusual place because it is one of the diagnosed strategies to discover the patron churn, however it isn't always suitable for complex problems [1]. But the observe shows that decreasing the information improves the accuracy of the selection tree [2].In a few cases, information mining algorithms are used for patron prediction and historical analysis. The strategies of regression bushes were discussed with other commonly used information mining strategies like selection bushes, rule-based learning, and neural networks [3].

### *Proposed System*

In this system, we use various algorithms like, XgBoost Classifier & Logistic Regression to find accurate values and which enables us to expect the churn of the customer. Here we implement the version by having a dataset that is trained and tested, which makes us have most accurate values. Fig.1 shows the proposed version for churn prediction and describes its steps. In the Initial step, facts preprocessing is performed in which we do filtering facts and convert facts into a similar form, and then we make feature selection.

Most of the literature focused more on data mining algorithms, but only a few of them focus on distinguishing the important input variables for churn prediction to be used for data mining algorithms implementation. Additionally, only noticeably one literature that had actually combined social network based variables in the input variables for data mining algorithms implementation. Moreover, the class imbalance problem was found to be not addressed on some of the literature.

In this Research we use Logistic Regression algorithm and Xgboost classifier to calculate churn Rate and to predict which factor affecting to customer.

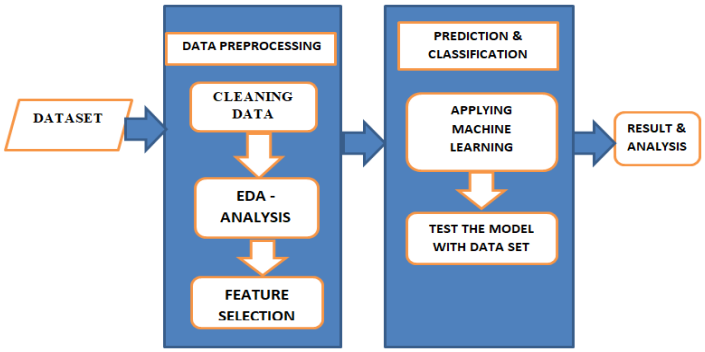


**Figure .1 Proposed System**

We use some algorithm and technologies for implement this model like Logistic Regression algorithm and Xgboost classifier.

In the further step prediction and classification is done using the algorithms lik XGBoost, and Logistic Regression (LR). Training and checking out the version with the facts set, we observe the behavior of the customer and analyze them. In the very last step, we do evaluation primarily based totally at the effects obtained and expect the customer churn.

### A. *Data Set*

## II. METHODOLOGY

### B. *Data Preprocessing*

Data set is a collection of featuers and N number of rows.Many values are in different formats. In a dataset, they may be

As we know, the information set is the start line for everything; it have to have full-fledged information to make the system

research about the hassle. Datasets may be generated or advanced from the scrap facts to be had on the internet. Some problems we have to create a dataset that makes feel that tells how to reply primarily based totally on real-time inputs for the hassle datasets can be gathered from the internet each day. A dataset is a collection of information. Most commonly, a information set has contents of a unmarried database table, or a unmarried statistical information matrix, where each column of the table describes a specific variable, and every row fits a given member of the information set in question. The information set lists the values of the variables, such as height, the weight of an object, for every member of the information set. Each value is recognized as a datum. As we know, the information set is the start line for this process.

We have Data IBM Watson Telecom customer churn Dataset https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/ The name of the Data set is WA_Fn UseC_ Telco Customer Churn.csv. This data set contains 7043 rows and 21 columns. Currently the dataset does not seem to have an imbalanced dataset.

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| customer | gender | SeniorCiti | Partner | Dependen | tenure | PhoneSer | MultipleLi | InternetSe | OnlineSec | OnlineBac | DevicePrc | TechSupp | Streaming | Streaming | Contract | Paperless | Payment | Monthly C | TotalChar | Churn |
| 7590-VHV | Female | 0 | Yes | No | 1 | No | No phone | DSL | No | Yes | No | No | No | No | Month-to | Yes | Electronic | 29.85 | 29.85 | No |
| 5575-GNV | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | No | Yes | No | No | No | One year | No | Mailed ch | 56.95 | 1889.5 | No |
| 3668-QPY | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | Yes | No | No | No | No | Month-to | Yes | Mailed ch | 53.85 | 108.15 | Yes |
| 7795-CFO | Male | 0 | No | No | 45 | No | No phone | DSL | Yes | No | Yes | Yes | No | No | One year | No | Bank tran | 42.3 | 1840.75 | No |
| 9237-HQI | Female | 0 | No | No | 2 | Yes | No | Fiber opti | No | No | No | No | No | No | Month-to | Yes | Electronic | 70.7 | 151.65 | Yes |
| 9305-CDS | Female | 0 | No | No | 8 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | Month-to | Yes | Electronic | 99.65 | 820.5 | Yes |
| 1452-KIO | Male | 0 | No | Yes | 22 | Yes | Yes | Fiber opti | No | Yes | No | No | Yes | No | Month-to | Yes | Credit can | 89.1 | 1949.4 | No |
| 6713-OKO | Female | 0 | No | No | 10 | No | No phone | DSL | Yes | No | No | No | No | No | Month-to | No | Mailed ch | 29.75 | 301.9 | No |
| 7892-POO | Female | 0 | Yes | No | 28 | Yes | Yes | Fiber opti | No | No | Yes | Yes | Yes | Yes | Month-to | Yes | Electronic | 104.8 | 3046.05 | Yes |
| 6388-TAB | Male | 0 | No | Yes | 62 | Yes | No | DSL | Yes | Yes | No | No | No | No | One year | No | Bank tran | 56.15 | 3487.95 | No |
| 9763-GRS | Male | 0 | Yes | Yes | 13 | Yes | No | DSL | Yes | No | No | No | No | No | Month-to | Yes | Mailed ch | 49.95 | 587.45 | No |
| 7469-LKB | Male | 0 | No | No | 16 | Yes | No | No | No interne | No | No interne | No | No interne | No | No interne | Two year | No | Credit can | 18.95 | 326.8 | No |
| 8091-TTV | Male | 0 | Yes | No | 58 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | One year | No | Credit can | 100.35 | 5681.1 | No |
| 0280-XJG | Male | 0 | No | No | 49 | Yes | Yes | Fiber opti | No | Yes | Yes | No | Yes | Yes | Month-to | Yes | Bank tran | 103.7 | 5036.3 | Yes |
| 5129-JLP | Male | 0 | No | No | 25 | Yes | No | Fiber opti | Yes | No | Yes | Yes | Yes | Yes | Month-to | Yes | Electronic | 105.5 | 2686.05 | No |
| 3655-SNQ | Female | 0 | Yes | Yes | 69 | Yes | Yes | Fiber opti | Yes | Yes | Yes | Yes | Yes | Yes | Two year | No | Credit can | 113.25 | 7895.15 | No |
| 8191-XWS | Female | 0 | No | No | 52 | Yes | No | No | No interne | No | No interne | No | No interne | No | No interne | One year | No | Mailed ch | 20.65 | 1022.95 | No |
| 9959-WOF | Male | 0 | No | Yes | 71 | Yes | Yes | Fiber opti | Yes | No | Yes | No | Yes | Yes | Two year | No | Bank tran | 106.7 | 7382.25 | No |
| 4190-MFU | Female | 0 | Yes | Yes | 10 | Yes | No | DSL | No | No | Yes | Yes | No | No | Month-to | No | Credit can | 55.2 | 528.35 | Yes |
| 4183-MYF | Female | 0 | No | No | 21 | Yes | No | Fiber opti | No | Yes | Yes | No | No | Yes | Month-to | Yes | Electronic | 90.05 | 1862.9 | No |
| 8779-QRD | Male | 1 | No | No | 1 | No | No phone | DSL | No | No | Yes | No | No | Yes | Month-to | Yes | Electronic | 39.65 | 39.65 | Yes |
| 1680-VDC | Male | 0 | Yes | No | 12 | Yes | No | No | No interne | No | No interne | No | No interne | No | No interne | One year | No | Bank tran | 19.8 | 202.25 | No |
| 1066-JKS | Male | 0 | No | No | 1 | Yes | No | No | No interne | No | No interne | No | No interne | No | No interne | Month-to | No | Mailed ch | 20.15 | 20.15 | Yes |
| 3638-WEA | Female | 0 | Yes | No | 58 | Yes | Yes | DSL | No | Yes | No | Yes | No | No | Two year | Yes | Credit can | 59.9 | 3505.1 | No |

**Figure.1 Data Set**

Duplicate values or null values that may lead to some loss inaccuracy, and there may be dependent.

Data have been collected from different sources, so there use adifferent type of format to notate a single value like gender someone represents M/F or Male/Female. The machine can understand only 0 and 1, so an image will be in 3-dimension data should be reduced to a 2-dimension format like data showto free from noisy data, null values, an incorrect size. Data cleaning can be performed by panda's tabular data and Open CV for images.

### *1. Data Filtering and Noise Removal*

It is very crucial to make the data useful because unwanted ornull values can cause unsatisfactory results or may lead to producing less accurate results. In the data set, there are a lot ofincorrect values and missing values. We analyzed the whole dataset and listed out only the useful features. The listing of features can result in better accuracy and contains only valuable features.

### *2. Feature selection & Engineering*

Feature selection is a crucial step for selecting the requiredelements from the data set based on the knowledge.

The dataset used here consists of many features out of whichwe chose the needed features, which enable us to improve performance measurement and are useful for decision-makingpurposes while remaining will have less importance. The performance of classification increases if the dataset is havingonly valuable variables and which are highly predictable. Thus having only significant features and reducing the number of irrelevant attributes increases the performance of classification.

### *3. Prediction & Classification*

Many techniques have been proposed for customer churn prediction in the telecommunication industry. In these three modeling techniques are used as predictors for the churn prediction. These techniques are outlined as

***1) Logistic Regression***

By using logistic regression, we can predict the probability of a churn i.e., the likelihood of a customer to cancel the subscription. Logistic regression is a supervised learning algorithm used for classification. In Logistic regression, we seta threshold; based on the limit, and only the classification is made using logistic regression. The threshold value is variable, and it is dependent on the classification problem itself.

Logistic Regression is primarily used for classification problems. The most common examples would be classifying whether an email is spam or not spam, and in this case, classifying whether a customer may churn or no. Logistic Regression makes use of a sigmoid function to map the data into two categories, of 0 and 1 in terms of a binary classification. Logistic Regression makes use of Maximum Likelihood Regression to estimation of regression coefficients. The coefficients obtained are that set of coefficients for which the probability of getting the data we have observed is maximum. This is used as the benchmark model. Logistic Regression was chosen here as it is one of the most common regression algorithms for binary classification in supervised learning.

***2) XGBoost***

XGBoost is the abbreviation for eXtreme Gradient Boosting. The number one cause of the use of XGBoost is because of its execution speed, and it's model performance. XGBoost uses ensemble learning methods; i.e., it uses a combination of different algorithms and produces output as a unmarried model. XGBoost helps parallel and disbursed computing while imparting efficient memory usage.

- Gradient Boosting is a method where new models are created to rectify the residual errors of the previous learner to ultimately make a strong learner. It is called gradient boosting because it uses gradient descent to minimize the loss function.
- XGBoost algorithm is an implementation of gradient boosted decision trees for performance and speed.
- ***This model is well suited for structured and tabular data and this was used to beat the performance of Logistic Regression.***

II .PROPOSED WORK

Initially, we will get the dataset from IBM, and by data filtering, we removed all the null values. Then we converted allthe data into a similar form, which more natural to understandand analyze. By Using EDA analysis we Explore the Data And Find Which Mainly Features Effect to Churn By using Logistic regression and having a different approach, we try to implement a predictor model for the Telecom company. Here we have a customer data set, and by preprocessing and feature selection, we divide the data set for training and testing. For this algorithm, we have made some feature engineering to have more efficient and accurate resultsusing that algorithm.

Logistic regression helps us to have a discriminative probabilistic classification and can estimate the probability of occurring event places. The dependent variable presents the event occurrence By training, the data to that model will get a result having their details, and then we will test the model with the remaining amount of data. Therefore we will get an accuracy based on the findings by which we can predict the customer.

Logistic Regression has shown an accuracy of around 78.60sss percent, which was sent as the benchmark metric to beat.

Similarly, we used the other two techniques to know which will provide us more accurate results. In the Random Forest, we used the same dataset, and by applying the technique, we trained the model and tested it out to get the results in the confusion matrix, which will show us the obtained output, andwe can notice the accuracy (fig.3). The result obtained from theXGBoost model is shown in (fig.4), where we can observe theaccuracy obtained by using that technique.

***III. RESULT AND ANALYSIS***

• The final parameters for XGBOOST were chosen because they gave the best accuracy, of around 80.5 percent. This was obtained by tuning each of the parameters and using grid search for each of the parameters.

• However, when a new model was created, its accuracy was

only slightly better than that given by a logistic regression

| Model | Accuracy(in percent) |
|---|---|
| Logistic Regression | 78.89 |
| XGBOOST | 80.05s |

**CONCLUSIONS**

The importance of churn prediction will help many companies, particularly in telecom industries, to have a profitable earnings and obtain desirable revenue. Customer churn prediction is the fundamental difficulty withi nside the Telecom Industry, and because of this, companies are looking to preserve the prevailing ones from leaving rather than acquiring a new patron. Three tree-based algorithms were chosen because of their applicability and range in this sort of application. By the use of XGBoost, and Logistic regression, we are able to get greater accuracy evaluating different algorithms. Here we're the use of the dataset of a few customers about their service plan and checking the values of them and have a precise prediction, which will help to identify the customers who are going to migrate to different corporation services. By this, the Telecom Company could have a clean view and can offer them a few exiting gives to live in that service. The obtained effects show that our proposed churn model produced higher effects and achieved higher via way of means of the use of system getting to know strategies. Random Forest produced higher accuracy among the various methods.

In the coming days, we will similarly research on Neural Network getting to know strategies to have higher patron churn prediction. To know the changing behavior of the customers, the study can be extended via way of means of the use of Artificial Intelligence strategies for trend analysis and patron prediction.

By exploring data, we came to know that features like Tenure, Monthly charges, Usage, Contract types, Payment methods are helpful to capture Churn customers.

performed. The data was split into training and testing sets. Logistic Regression was chosen as a benchmark model whose performance XGBoost and amartificial neural network is supposed to beat.

**REFERENCES**

***PAPERS***

[1] Kuanchin Chen, Ya - Han Hu, Yi - Cheng Hsieh, "Predicting CustomerChurn from Valuable B2B Customers in the Logistics Industry: A CCase Study", Information Systems and e-Business Management, Volume 13 Issue 3, August 2015, Pages 475-494.

[2] Junxiang Lu, "Predicting Customer Churn in the Telecommunications Industry -- An Application of Survival Analysis Modeling Using SAS", In SAS Proceedings, SUGI27, pages 114-127, 2002.

[3] Khalida binti Oseman, Sunarti binti Mohd Shukor, Norazrina Abu Haris,Faizin bin Abu Bakar, "Data Mining in Churn Analysis Model for Telecommunication Industry", Journal of Statistical Modeling and Analytics, Vol.1 No. 19-27, 2010.

[4] Shan Jin, Yun Meng, Chunfen Fan, Feng Peng, Qingzhang Chen, "The Research on Applying Data Mining to Telecom Churn Management", Lecure Notes in Information Technology - Proceedings of 2012 2ndInternational Conference on Materials, Mechatronics and Automation (ICMMA 2012), 2012.

[5] Vadakattu, B. Panda, S. Narayan, and H. Godhia, ‘‘Enterprise subscription churn prediction,’’ in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 1317–1321.

[6] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, ‘‘Telecommunication subscribers’ churn prediction model using machine learning,’’ in Proc. 8th Int. Conf. Digit. Inf. Manage., Sep. 2013,pp. 131–136

[7] S. Babu1, Dr. N. R. Ananthanarayanan2 Predicting Churn Customer in Telecom using Peer grading Regression Learning Technique www.ijser.in (2014)

[8]A Survey on Customer Churn Prediction using Machine Learning Techniques Saran Kumar A. M.E. Scholar Kumaraguru College of Technology Coimbatore, India Chandrakala D., PhD Professor Kumaraguru College of Technology Coimbatore, India International Journal of Computer Applications (0975 – 8887) November 2016

[9] Churn Analysis and Prediction A Case Study based-on Decision Tree and Neural Network in Logistics Sector (IJCSIS) International Journal of Computer Science and Information Security,

Vol. 14, No. 8, August 2016

[10] M.HemaLatha, S.Mahalakshmi Predicting Churn Customer in Telecom using Peergrading Regression Learning TechniqueInternational Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-6, April 2020

***WEBSITES***

[1] https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets

[2] https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn

[3] https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html

[4]https://www.smartkarrot.com/resources/blog/how-to-design-a-great-customer-churn-prediction-software-algorithm/

[5] https://towardsdatascience.com/predicting-customer-churn-using-logistic-regression-c6076f37eaca