International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-3, May 2022
https://doi.org/10.55524/ijircst.2022.10.3.10
Article ID IRP1266, Pages 54-57
www.ijircst.org

Customer Churn Prediction in Telecom Industry Using Regression Algorithms

P. Geetha Priyanka[1], and Mr. Sk. Althaf Rahaman[2]

[1]Student, Department of Computer Science, GITAM Deemed to be University, Vishakhapatnam, India
[2] Assistant Professor, Department of Computer Science, GITAM Deemed to be University, Vishakhapatnam, India

Correspondence should be addressed to P. Geetha Priyanka; geetha.paruchuri02@gmail.com

ABSTRACT- Customer acquisition and retention is a major challenge in a variety of industries, but it is most severe in highly competitive and fast-growing companies. Customer turnover is a major worry for large organisations since keeping a loyal customer is significantly more valuable than gaining a new one. Finding the causes that cause customer turnover is critical for implementing the appropriate solutions to prevent and reduce churn. The goal of this study is to employ machine learning (ML) algorithms to detect prospective churn clients, categorise them based on usage patterns, and illustrate the findings of the analysis. Extra Trees Classifier, XGBoosting Algorithm, and Decision Tree, Random Forest have the best churn modelling performance, especially for 80:20 dataset distribution, with AUC scores of 0.85, 0.96, and 0.977, respectively.

KEYWORDS- Machine Learning, Logistic Regression, Churn Prediction, Feature Engineering, and Accuracy Score.

I. INTRODUCTION

Any firm can't afford to lose clients. According to the authors of "Leading on the Edge of Chaos" 2% reduction in customer attrition equates to a 10% cost reduction [1] . Furthermore, according to the White House Office of Consumer Affairs gaining new customers is 6–7 times more expensive than keeping existing ones. Early detection of dissatisfied consumers, taking into account their values and the danger of churn, allows you to provide them incentives to stay. Cancellation of a subscription, account closure, non-renewal of a contract or service agreement, or switching to another service provider are all examples of customer churn.

Churn can happen for a variety of reasons, and customer churn analysis can help determine the source and timing of churn so that effective churn retention efforts can be implemented [2].

The goal of customer churn prediction is to identify customers who are likely to depart a business. In order to keep existing clients, the banking industry must understand the causes of churn, which can be discovered via data-driven knowledge. When a corporation takes a reactive approach, it waits until the consumer asks for their service relationship to be terminated. In this case, the corporation will provide an incentive for the consumer to stay. When a business takes a proactive approach, it looks for consumers who are likely to leave before they do. The corporation then offers unique incentives to these clients in order to keep them from leaving.

II. FRAMEWORK AND METHODOLOGY

A. Methodology

The Process of analysis is shown in figure 1, It shows the process going in the Prediction.

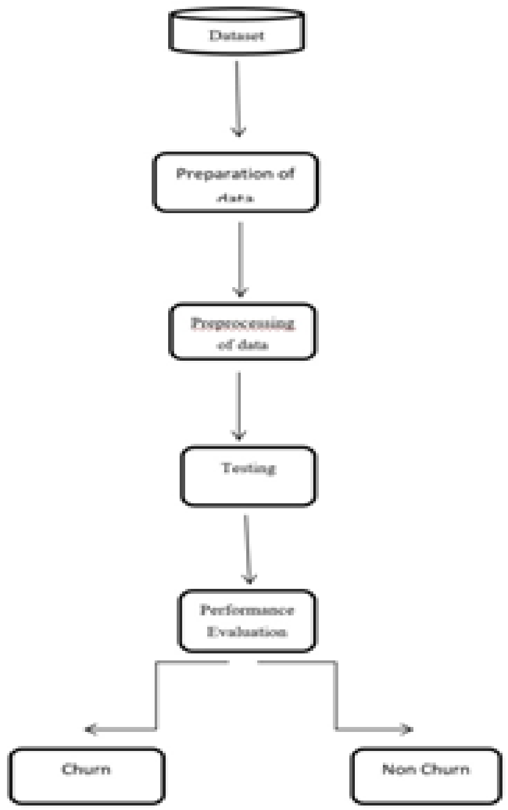

Figure. 1: Process of Analysis

B. Dataset

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation [3]. Then select the "Equation" mark-up style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents [4]. Use parentheses to avoid ambiguities in denomina we use a dataset file with .csv extension that is publicly available on Kaggle. And the dataset has different Attributes with numeric and string contents.

1- State: the two-letter acronym for the US state in which the customer resides.

2- Account Length: the amount of time this account has been active.

3- Area Code: the three-digit area code associated with the customer's phone number.

4- Phone: the last seven digits of your phone number

5- International Plan: the customer's international calling plan: yes/no

6- VMail Plan: whether or not the consumer has voice mail: yes/no

7- VMail Message: apparently the monthly average for voice mail messages.

8-Day Mins: The total amount of calling minutes used during the day.

9-Day Calls: the total number of calls made in a given day.

10-day period: The billed cost of daytime calls for a.

11- Eve Mins: total number of minutes

12- Eve Calls: the total number of calls made that night.

13- Eve Charge: the cost of calls made in the night

14- Night Minutes: the total number of calling minutes used at night.

15- Night Calls: the total number of calls made at night.

16- Night Charge: the fee of calls made at night.

17- International Minutes: total international minutes

18- International Calls: total international calls

19- International Charge: the charge for international calls.

20 : the total number of calls to Customer Service.

21- Churn? : Did the consumer leave the service? True/false

C. Pre-processing Data

Data cleaning, feature selection, and data transformation are all phases in the pre-processing of data. Each stage is described in detail below:

Data transformation consists of two explanatory variables that can be converted from binomial to binary form to make the models more relevant[4].

Missing data imputation or handling is part of the data cleansing process. Because some of the algorithms employed cannot handle missing data, missing values can be translated using median, mean, or zero[5]. However, replacing missing data with a statistically derived value is a superior option. There are missing values in two categorical variables and some numerical variables in the data set.

One of the most important aspects that can influence the model's performance before it is trained is feature selection.

.D. Data Cleaning

The practise of correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning.

E. Feature Selection

The process of selecting the features that contribute the most to predict variable or output. Advantages of feature selection include:

- Enhance your performance
- Enhances precision

III. CLASSIFICATION AND ALGORITHMS

The process of converting raw data into features that can be utilised to build a prediction model with machine learning. When compared to sending merely raw data to a machine learning process, the motivation is to leverage these extra features to improve the quality of results from a machine learning process.Numpy, Pandas, Scikit, and Matplotlib were used to analyse the data [6]. Scikit is a simple data processing and analysis toolbox. The majority of the work is done with these tool sets.

The problem has been defined as predicting neither or not a customer will churn. As a result, it's a classification issue [7]. The data must be separated into three random sets before our learning system can be trained. For instance:

- Approximately 60% of the data will be utilised for training different algorithms or variations on the same method.
- Another 20% of the data is used in cross-validation to find the model with the fewest mistakes.
- The remaining 20% will be used to test the winner model's accuracy using cross-validation.

Predicting which customers will leave is only the beginning. In Machine learning, we use supervised learning algorithms like as follows:

- Logistic Regression
- Random Forest
- XGBoost Classifier

IV. COMPARISION OF ALGORITHMS

Table 1 gives comparisons between the algorithms with the accuracy

ALGORITHMS	TRAIN ACCURACY	TEST ACCURACY
Logistic Regression	0.866018198933166	0.8541862652869238
Random Forest	0.964541158468749	0.9774223894637818
XGBoost	0.9996862252902416	0.9774223894637818

Table 1: Accuracy Table

A. Evaluation

Results Interpretation: After evaluating if our selected model accurately forecasts customer attrition, it's a good sign that our hypotheses are valid. If not, more data analysis and questions to persons with domain knowledge will be required to establish alternative theories, reformat certain features, or create new ones [8]. However, we must train and test our models in order to determine the impact of these adjustments [9].

V. RESULTS

This initiative included several measures. The proposed solution makes use of a number of Python classification methods. These are the most common Machine Learning strategies for getting the most accuracy out of data. In this investigation, we can see that the XGBOOSTclassifier outperforms the others. Overall, we used the most advanced Machine Learning techniques to forecast performance and achieved high accuracy. Figure 2 shows the accuracy of the algorithms.

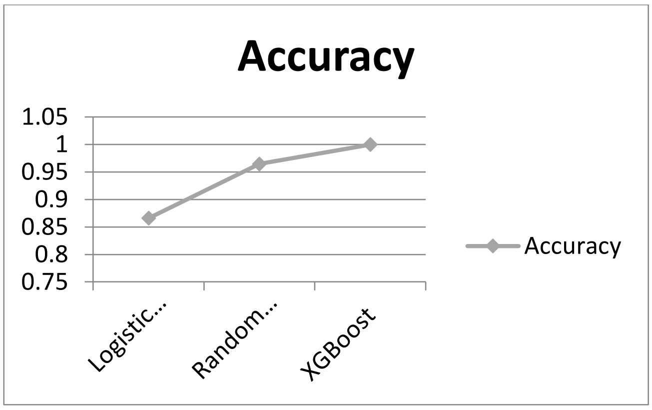

Figure 2: Accuracy Graph

VI. CONCLUSION

Understanding, data collecting and processing, hypotheses development, modelling, results evaluation, and reporting to model deployment is a critical conceptual tool for thinking about data science projects like our design to anticipate customer attrition. A successful data science project requires a thoughtful compromise between the data's capabilities and the project's objectives. Churn prediction, or the effort of identifying consumers who are likely to stop using a service, is a lucrative and important issue in the telecom industry. Customer churn is a major issue in the telecom industry because customers do not hesitate to quit if they do not know what they are looking for. Customers want value for money, lower prices, and higher service quality. Customer happiness is directly linked to customer churning. Customer acquisition costs are higher than client retention costs, making customer retention a tough business model to model. There is no conventional approach that accurately tackles the churning difficulties of global telecom service providers.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] C. Blank and T. Hermansson, "A Machine Learning approach to churn prediction in a subscriptionbased service," KTH, Stockholm, 2018.

[2] D. Buö and M. Kjellander, "Predicting Customer Churn at a Swedish CRM-system Company," Linköpings Universitet, Linköping, 2014.

[3] K. Mishra and R. Rani, "An inclusive survey on machine learning for CRM: a paradigm shift," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017.

[4] H. Gebert, M. Geib, L. Kolbe and W. Brenner, "Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts," Journal of Knowledge Management, vol. 7, no. 5, pp. 107-123, 2003.

[5] M. Sergue, "Customer Churn Analysis and Prediction using Machine Learning for a B2B SaaS company," KTH, Stockholm, 2020.

[6] F. Khodakarami and Y. Chan, "Exploring the role of customer relationship management A. (CRM) systems in customer knowledge creation," Information & Management, vol. 51, pp.27-42, 2014.

[7] D. L. Garcia, A. Nebot and A. Vellido, "Intelligent data analysis approaches to churn as a business problem: a survey," Knowledge and Information Systems, vol. 51, no. 3, pp. 1-56, 2017.

[8] N. Gordini and V. Veglio, "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry," Industrial Marketing Management , vol. 62, pp. 100-107, 2017.

[9] Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A churn prediction model using random forest"

ABOUT THE AUTHORS

P. Geetha Priyanka, Student, Department of Computer Science, GITAM Deemed to be University, Vishkhapatnam, Andhra Pradesh, India.

Mr. Sk. Althaf Rahaman, M.Tech (Ph.D) Assistant Professor, Department of Computer Science, GITAM Deemed to be University, Vishakhapatnam, Andhra Pradesh, India. His Research areas are Big Data Analytics, Machine Learning, and Artificial intelligence.