ISSN 2278-3091
Volume 10, No.3, May - June 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1501032021.pdf
https://doi.org/10.30534/ijatcse/2021/1511032021

Software Development and Modelling for Churn Prediction Using Logistic Regression in Telecommunication Industry

Syed Zain Mir[1], Azfar Ghani[2], Sajid Yasin[3], AzeemAftab[4], Ikram e Khuda[5]
[1]Iqra University, Pakistan, szainmir@iqra.edu.pk
[2]Iqra University, Pakistan, azfarghani@iqra.edu.pk
[3]Iqra University, Pakistan, sajidyasin@iqra.edu.pk
[4]Iqra University, Pakistan, azeem.cheema@iqra.edu.pk
[5]Iqra University, Pakistan, ikram@iqra.edu.pk

ABSTRACT

Along with the fast progress of the telecom industry, a good and reliable customer relationship has likely become the main concern for the telecom service providers. It is known that if a standing customer dismisses a bond with current wireless company and avail the services of another wireless company results the loss of customer which is referred as churn customer. All telecommunication service providers are affected badly from deliberate churn. The survival of these companies depends on its ability to hold customers. This paper focus to identify the best modelling technique which helps to correctly predicts the churn customer and also emphasis to make a reliable software for the telecom companies to find which customer is going to churn, java programming is done in eclipse neon version for software application and logistic regression technique is used to make a mathematical model, because most of the statistician believes that when the independent variable in a dataset does not distributed normally, logistic regression is a best suited and acceptable modelling technique than other modelling techniques.

Key words: Customer churn, constraints, retention, sensitivity, specificity, telecom industry.

1. INTRODUCTION

Telecom corporations measure intentionally churn customer by a monthly figure of 1.9% to 2.1%. The average churn rate per annum is measured 10% to 67%. Customer churn prediction model enables the service provider to detect the reasons of customer churn. If the customers were pleased with their current service provider, then the treatment towards customer and its services would result that they would not be looking around for other networks[4].An analysis of customer retention shows that the most companies lost their customers due to dissatisfactory treatment towards their customers[3].

The purpose of this research is to find the parameters which involves significantly in customer churn and design a mathematical model using logistic regression analysis technique for churn prediction and to further support this work software application has also been developed to predict the churn customer by using certain parameters for telecom companies so that this could be utilize by the industry[5].In this paper, online purchase customer's behavior analyzed, using decision tree induction method, customers are divided into groups or categories namely premium, best, and moderate or churn. Correlation and linear least square regression was carried out to predict the customer purchasing behavior [18].This study assists the telecom companies to enhance their customer retention and reduce the risk of customer churn hazard by increasing the predicting power for customer churn[2]. The outcome of this analysis helps the wireless telecom service provider to manage their resources efficiently to improve the customer satisfaction which can help to reduce the churn customers[13].

2. RESEARCH METHODOLOGY

The important aspect of the study was to get a reliable dataset which could be used for mathematical modeling. After getting the widely used dataset by the statistician, logistic regression algorithm applied on it for attainment of a mathematical model for churn prediction. There were initially twenty independent variables and one dependent variable. Mathematical model comprises all these independent variables has the higher computational load as well as increase the complexity of the model. To reduce the complexity, it was required to use only significant variables which could have the higher impact on churn prediction, for this purpose iterative method has been adopted and several iterations have been done.

This authorize to reduce the input variable from twenty to five variables and the computational load also been reduced. After designing a mathematical model, software application would need to be developed so that this work could be utilized by telecom industries. The software application has been created using java programming in eclipse neon version platform.

3. DATASET

The dataset of wireless telecom company is used which is available online for research purpose and can be downloaded from the GitHub Inc. website[1].This dataset is widely used by the researchers and statistician for mathematical modelling using different statistical analysis methods. The dataset comprises the data of last three months of 3333customers in which 477 customers are the churn customers. For every observation, there are Twenty-one features available and no missing values.

4. INPUT SELECTION AND FEATURES CONSTRAINS

The class variable is highly skewed in the available dataset, the percentage of churn customers is 14.3% while the percentage of customer who retained with current service provider is 85.7%[8] which effects the statistical modeling and cause complications in predicting the churn customers. To minimize the computational load, the logistic regression algorithm is applied gradually to exclude all non-significant variables in a dataset through several iterations. It is necessary to ensure that the dataset is comprised only of significant variables and by observing the accuracy of results (percentage of correctly predicted instances categorized as churn and no churn), sensitivity (percentage of correctly predicted churn customer) and specificity (percentage of correctly predicted retained customers), in each iteration, it was found that these are the following input variable as shown in Table 1 possess the higher impact on customer churn prediction.

Table 1: Customer Churn prediction impact

Feature Constraints	Description
Total Day Call Charges	Daytime call Charges per month ($/month)
Customer Service Calls	No. of calls to customer helpline
International Plan	Subscriber of International plan (0=no subscription, 1= subscriber)
Total International Call Charges	International Call charges per month ($/month)
Total Evening Call Charges	Charge for evening usage ($/month)

5. LOGISTIC REGRESSION ANALYSIS

This technique widely used in data mining, machine learning and can also be used traditional statistics[17]. Logistic regression is used for prediction when the dependent variable is categorical i.e. some event is happening or not happening. In classification problems, this method has an edge and can compete with other modelling techniques.

The logistic regression analysis can be interpreted by the concept of odds ratio which is the ratio of the probability of some event happening *p* or not happening *q which is(1-p).*

$$\text{Odds } (\theta)=\frac{p}{1-p} \qquad (1)$$

The logistic regression modelling is used to estimate the value of *p* which is represented as $\hat{p}$.

The aim is to estimate the value of *p* using Logistic regression analysis according to the classification of output variables which is churn and no churn in our dataset.

The prediction of the target variable is transform by the nonlinear function called the logistic function. Graphically, it is a S shaped curve and known as Sigmoid function can map any real value between 0 and 1. Statistician developed the logistic function to evaluates the different properties of a training dataset

$$g(z) = \frac{1}{1-e^{-z}} \qquad (2)$$

consider *z* is the linear function of multivariate regression model then the equation for z is:

$$z = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + \ldots + \beta n\ Xn \qquad (3)$$

The prediction of the classification variable by logistic regression analysis would found by following equation and by setting up the threshold value output will be decided accordingly either churn or no churn:

$$\hat{p} = \frac{1}{1-e^{-(\beta 0 + \beta 1\ X1 + \ldots + \beta nXn)}} \qquad (4)$$

Where,

$\hat{p}$is the estimated output

$X_i$is the predictors

$\beta_0$is the constant

βi is the weight of each predictor

6. CONFUSION MATRIX

Confusion matrix evaluates the performance of Binary Classification model[6]. This is the matrix of the models predicted class versus the actual class as shown in Table II. A confusion matrix has two columns and two rows having information about the number of true positives, false positives, false positives, true negatives[10].

The detailed analysis can be done by sheer proportion of correct classification. Confusion matrix is easy to understand, though the relative terms might be confusing. The performance of the mathematical model can be evaluating by finding the sensitivity, specificity and classification accuracy[7].

		Actual	
		yes	No
Predicted	Yes	TPs	FPs
	No	FNs	TNs

Table 2: Matrix of the Model

Where;

- *TPs* is the value of True Negatives
- *FN*s is the value of False Negatives
- *TN*s is the value of True Negatives and
- *FPs* is the value of False Positives

7. RESULT AND DISCUSSION

The classification table of the proposed mathematical model is based on the classification threshold of 0.1485 in Table 3,indicates the accuracy of predicted positive is 75.3623% and for predicted negative the accuracy is 76.1404% while the classification accuracy of the model is 76.0276%.

Table 3: Proposed Model Accuracy

	Actual Chun	**Actual No churn**	**Correct Prediction Percentage**
Predicted Churn	364	680	75.3623
Predicted No churn	119	2170	76.1404

The linear function of the logistic regression is:

$$z = -7.56205 + 0.073955\mathbf{x_1} + 0.498887\mathbf{x_2} + 1.934837\mathbf{x_3} + 0.295802\mathbf{x_4} + 0.076536\mathbf{x_5} \quad (5)$$

Equation for the prediction of target variable is:

$$\hat{p} = \frac{1}{1-e^{-z}} \quad (6)$$

7.1 ROCCurve:

ROC (Receiver operating Characteristics) curve is the graphical representation of the performance of classification model [14],[8]. The ROC curve is the relation between rate of false positive (i.e. sensitivity) on x-axis and rate of true positive on y-axis. The accuracy of the model can be determining by finding the area under the curve. Greater area represents the better performance of the model [11]. After several iterations, at a cut off value 0.1484 the proposed model depicts the best results and the highest area under the curve (AUC) as shown in Figure 1.

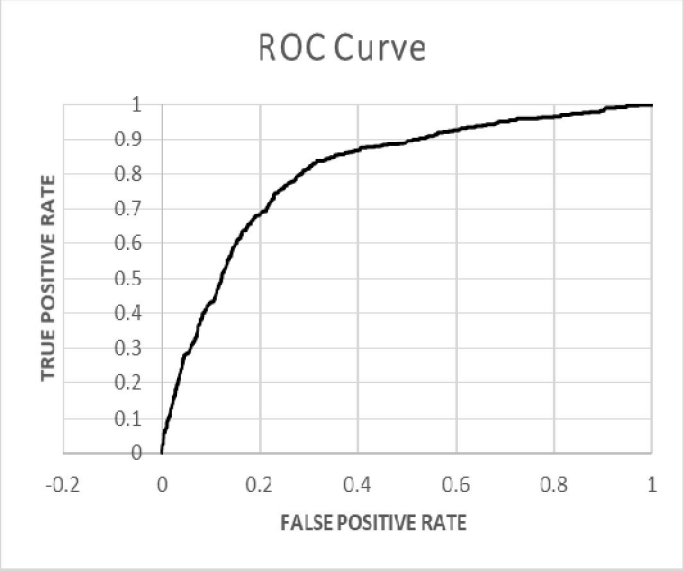

Figure 1: False Positive Rate Analysis

7.2 Classification Accuracy

Classification accuracy (CA) is one parameter for accessing the predicting model for the binary classification. Classification accuracy helps to know how well the classification model can predict the value of the predicted attribute[9]. The classification accuracy of the proposed binary classifier is 76.0276%, which is calculate by the formula given as follows:

$$CA=\frac{TNs+TPs}{TNs+TPs+FNs+FPs} \quad (7)$$

7.3 Sensitivity and Specificity

Sensitivity and specificity evaluates the accuracy of the model. Sensitivity is defined as the proportion of the customers who actually churn to the customers identified as churn[12]. Specificity is known as the fraction of the correctly identified non-churners. The following are the equations for the sensitivity and specificity shown.

$$\text{Sensitivity} = \frac{TPs}{TPs+FNs} \quad (8)$$

$$\text{Specificity} = \frac{TNs}{TNs+FPs} \quad (9)$$

Telecom service provider prefers high sensitivity model as compared to the model with high specificity because the wrong estimation of churners can affect the cost much higher than the wrong estimation of non-churners [15].The sensitivity and specificity of the proposed model is given in Table 4.

Table 4: Compared with high Sensitivity &high Specificity

Sensitivity	0.348659
Specificity	0.948012

8. SOFTWARE DEVELOPMENT

The lifecycle for software development is completely followed is given in Figure 2.

A Software was developed by java programming in eclipse neon version for the prediction whether the customer churn or not using the values of independent variable shown in Table I. The code was designed to fulfill the requirements of service providers, and after thoroughly testing and fixing several bug, the software is ready for deployment for the telecom industry.

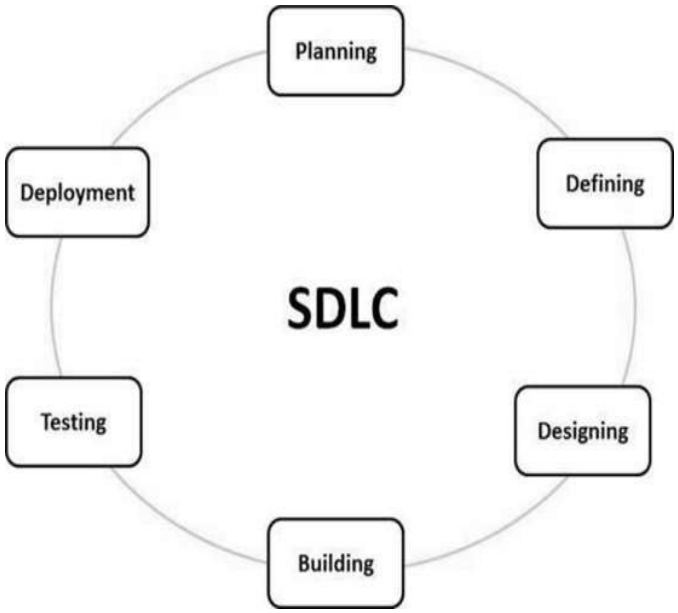

Figure 2: System Development Life Cycle

The Graphic User Interface(GUI) of the software is user friendly. It can be used simply by telecom service provider with no additional skills required[16]. In entering the five feature constraint for the last three months of any user, it can predict instantly whether the customer is expected to churn or not. The snapshot of the GUI of churn prediction software is shown in Figure 3.

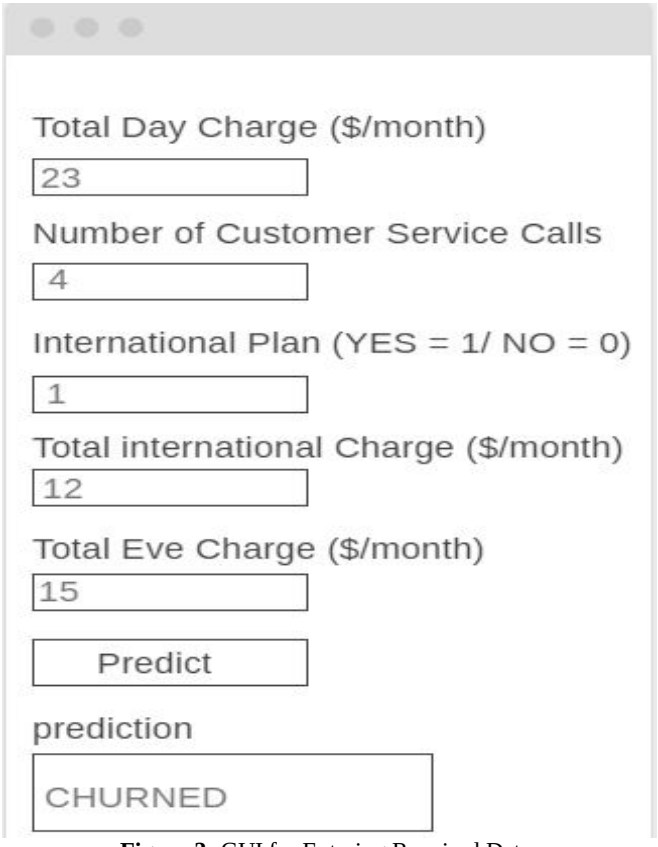

Figure 3: GUI for Entering Required Data

9. CONCLUSION

The effects of proposed model by using logistic regression analysis achieved a good accuracy which helps the telecom service provider to minimize the customer churn and manage the acceptable level. Telecommunication companies face complications to increase the customer retention rate and minimize an unavoidable loss in their business due to high churn rate. The churn prediction software also assists to boost up the predicting power. Simpler GUIs of predicting software and efficient modelling of telecommunication churning techniques can enhance the business and stop the telecom industry for upcoming challenges.

10. ACKNOWLEGEMENTS

I would like to thanks Sir Ikram e khuda for his guidance, suggestions and discussions during research and Sir Umer Chawla for his guidance in software development.

REFERENCES

1. GitHub Inc. website.**https://github.com/VarunRaosr/Telecom_churn_machine_learning/blob/master/bigml_59c28831336c6604c800002a.csv**, VarunRaosr/Telecom_churn_machine_learning.
2. A Idris, A Iftikhar, Z urRehman. **Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO under sampling,** Cluster Computing, Springer. 2017.
3. Vishal Mahajan, RichaMisra , Renuka Mahajan.**Review onfactors affecting customer churn in telecom sector,** Int. J. Data Analysis Techniques and Strategies, Vol. 9, No. 2. 2017.
4. B. E. A. Oghojafor, G. C. Mesike, C. I. Omoera and R. D. Bakare. **Modelling telecom customer attrition using logistic regression,** African Journal of Marketing Management Vol. 4, No. 3, pp. 110-117.2012.
5. Amin, Adnan, Saeed Shehzad, Changez Khan, Imtiaz Ali, and Sajid Anwar, **Churn prediction in telecommunication industry using rough set approach.** In New Trends in Computational Collective Intelligence, pp. 83-95. Springer, Cham, 2015.
6. Ikram E Khuda, Syed Zain Mir, Mansoor Ebrahim and Kamran Raza.**Simulation and Modeling of Fading Gain of the Rayleigh Faded Wireless Communication Channel Using Autocorrelation Function and Doppler Spread,**Journal of Applied Science and Engineering, Vol. 22, No. 4, pp. 637644 (2019).
7. Xie, Y., Li, X., Ngai, E.W.T. and Ying, W. **Customer churn prediction using improved balanced random forests,**Expert Systems with Applications, Vol. 36, No. 3, pp.5445–5449. 2009.
8. Kim, H. and Yoon, C.**Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Telephony Market,** Telecommunications Policy, 28,75 1-765. 2004.
9. Wouter Verbeke, David Martens, Christophe Mues, Dart Baesens. **Building comprehensible customer churn prediction models with advanced rule induction techniques,** Expert Systems with Applications, Vol. 38, pp.2354–2364. 2001.
10. Umman Tugba Simsek Gursoy.**Customer churn analysis in telecommunication sector,** Istanbul University Journal of the School of Business Administration, Vol. 39n No. 1, pp.35-49. 2010
11. John Hadden, Ashutosh Tiwari, Rajkumar Roy. **DymitrRuta, Computer Assisted Customer Churn Management: State-Of-The-Art and Future Trends,** Computers & Operations Research. Vol. 34, No. 10, pp. 2907-2917. 2007.
12. Mohammed Hassouna, Ali Tarhini, Tariq Elyas&Mohammad Saeed AbouTrab. **Customer Churn in Mobile Markets: A Comparison of Techniques,** International Business Research; Vol. 8, No. 6, pp. 224-227. 2015.
13. MR Khan, J Manoj, A Singh, J Blumenstock. **Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty,**IEEE International Congress on Big Data, 677-680. 2015.
14. W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens. **Newinsights into churn prediction in the telecommunication sector: A profit driven data mining approach,** Eur. J. Oper. Res., vol. 218, no. 1, pp.211–229, Apr. 2012.
15. C.-P. Wei and I.-T. Chiu. **Turning telecommunications call details to churn prediction: a data mining approach,** Expert Syst. Appl., vol. 23, no.2, pp. 103–112, Aug. 2002.
16. Verbraken, Thomas, WouterVerbeke, and Bart Baesens. **A Novel Profit Maximizing Metric for Measuring performance of customer churn prediction models,** IEEE Transaction on Knowledge and data Engineering, 2012.
17. J. Franklin. **The elements of statistical learning: data mining, inference and prediction,** Math. Intell., vol. 27, no. 2, pp. 83–85, Nov. 2008
18. Rahman, NurShamsiah Abdul, Lim Ken Tak, and AnisFarihan Mat Raffei. **Preliminary Studies on Predicting Customer Purchase Behaviour in Online Retail Business,** International Journal 9.1.4 (2020).