Churn Prediction using Machine Learning-An Analytical CRM Application

M Kavitha Margret, M Monishapriyadharshini, S Nathies, C Sriram

***Abstract:** CRM represents (Customer Relationship Management).It is a classification of programming that covers many arrangement of utilizations that are intended to support organizations and furthermore to oversee huge numbers of the business forms like client information. CRM framework models incorporate stages worked to oversee advertising, deals, client support, and backing, all associated with assistance organizations work all the more viably. With a CRM framework, organizations can dissect client collaborations and improve their client connections. The data based forecast models utilizing AI systems have increased monstrous prevalence during the most recent couple of decades. These models have been applied in enormous number of areas like clinical conclusion, wrongdoing expectation, films rating, and so forth. Thus it is utilized in telecom industry where models of expectation have been applied for the forecast of not fulfilled clients who are probably going to change the administrations and furthermore the specialist organization. In telecom the money related expense of client agitate is tremendous henceforth numerous organizations have examined different variables, (for example, cost of the call, nature of the call, client assistance reaction time and so on.) utilizing different AI strategies. This work proposes different ML strategies for client agitate expectation.*

***Keywords:** CRM, Telecom, churn, classification, Machine learning.*

I. INTRODUCTION

Current occasions are the time of rivalry among organizations and comparable is the pattern in correspondence industry. There is a tremendous challenge among the versatile administrators, because of which, the media transmission industry faces troubles to hold their present endorsers. Client stir anticipation is a significant piece of the Customer Relationship Management (CRM). Beat depicts the clients who have ended the relationship with their present specialist organization based on disappointment. The ongoing progression in examination of data frameworks techniques, especially being used of the data based forecast strategies, has additionally roused the media transmission industry. Telecom organizations put away immense framework and cash to actualize stir forecast models to locate the conceivable churners before their clients choose to change the specialist organization. Stir counteraction is a significant factor as the expense for client procurement is a lot more prominent than the expense of client maintenance. Subsequently, little upgrades in client maintenance can prompt significant benefit for the telecom organization. There is a requirement for exact agitate expectation models to distinguish the most probable churners and their motivations to beat. So as to forestall client stir in the telecom business, it is imperative to fabricate a viable client agitate expectation model. Various agitate expectation models have been proposed and actualized, which plan to distinguish clients with a high inclination to leave.

Revised Manuscript Received on March 11, 2020.

Ms. M. Kavitha Margret, Assistant Professor in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India.

Ms. M. Monishapriyadharshini, IV year CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs120@skct.edu.in

Mr. S. Nathies, IV year CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs127@skct.edu.in

Mr. C. Sriram, IV year CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs230@skct.edu.in

II. LITERATURE SURVEY

This methodology utilized/applied various ways to deal with foresee stir in the businesses of telecom. Practically each one of those methodologies utilized the idea of AI or information mining. The related work concentrated larger part on applying any one technique for either AI or information digging for information extraction while others associated with anticipating beat by looking at a few systems accessible. Gavril et al. displayed a technique for information mining which is progressed to anticipate stir for the most part for prepaid clients by utilizing dataset for call subtleties of 3333 clients which had 21 highlights, and one ward variable for beat expectation and this parameter has two qualities Yes/No. Here this methodology have certain highlights that incorporates data about approaching and active phone messages just as messages for every single client. The creator applied head part examination calculation "PCA" to lessen information measurements. Three AI calculations were utilized: Neural Networks, Support Vector Machine, and Bayes Networks to foresee stir factor. The creator utilized AUC to quantify the exhibition of the calculations. For Bayes Networks the AUC esteem is 99.10%. For neural systems the AUC esteem is 99.55% and 99.70% for help vector machine. Right now dataset utilized was little and no missing qualities existed. Neural system calculation to tackle the issue of client beat was proposed by He et al. This model is utilized for forecast in an enormous Chinese telecom organization that has about 5.23 million information of their clients. Here the forecast precision standard was the general exactness rate and it has arrived at 91.1%.

A methodology dependent on hereditary writing computer programs was proposed by Idris in broadcast communications with Ada Boost to display the stir issue. Two standard informational collections were utilized to test this model. One by cell2cell and the other by Orange Telecom with 89% precision for the cell 2 cell dataset and 63% for the Orange Telecom. Utilizing the large information stage Huang et al. contemplated the issue of client agitate.

The fundamental objective was to demonstrate that huge information will improve the way toward foreseeing the stir to extraordinary expand contingent upon the information assortment, volume and speed. The branch of Operation Support and Business Support office manages information at China's biggest broadcast communications organization. To design the cracks it needs a foundation of large information for additional procedure.

Here Random Forest calculation was utilized and assessed by utilizing AUC. In telecom Makhtar et al. utilized unpleasant set hypothesis and proposed a model for stir forecast. As referenced Rough Set arrangement calculation beat different calculations like Voted Perception Neural Network, Decision Tree and Linear Regression. The issue of uneven informational collections were contemplated by different specialists where the dynamic client classes are more prominent than the stirred client classes, as it is a significant issue in foreseeing beat. Six distinctive testing systems were analyzed by Amin et al. for oversampling telecom agitate forecast issue. The outcomes demonstrated that MTDF and rules-age dependent on hereditary calculations beat the other oversampling calculations that are looked at previously.

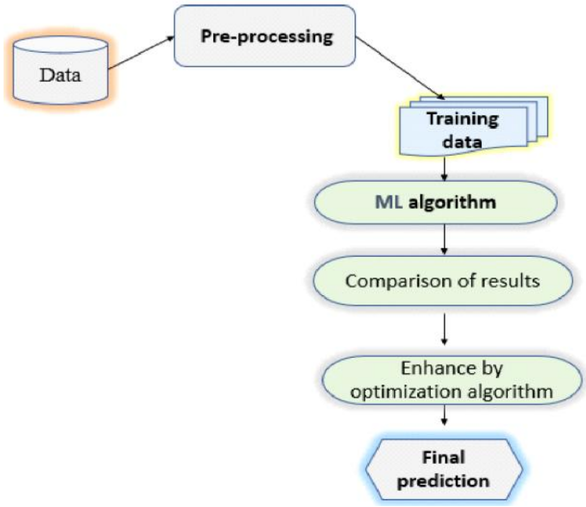

Fig. 1.Methodologies Invloved.

Figure 1 demonstrates the steps involved in predicting the churn in the given dataset using machine learning algorithms and getting the optimised churn value.

III. PREPROCESSING

LABEL ENCODING:

In machine learning, this methodology usually deals with datasets which contains multiple labels in one or more than one column.These type of labels can be represented in the form of letters ie.words or it can be numbers also. To make the data understandable or in human readable form, the training data is often labelled in words.The process of converting the labels into numerical form so that it will be converted into the machine-readable form is defined as Label Encoding.After that the labels can be processed/operated in a better way by the decisions of Machine learning algorithms. In supervised learning it is the most important pre-processing step for the structured datasets.Here Label encoding also assigns a unique number (starting from 0) to each class of the data. This may lead to the generation of priority issue in training of data sets. A label with high value may be considered to have high priority than a label having lower value.

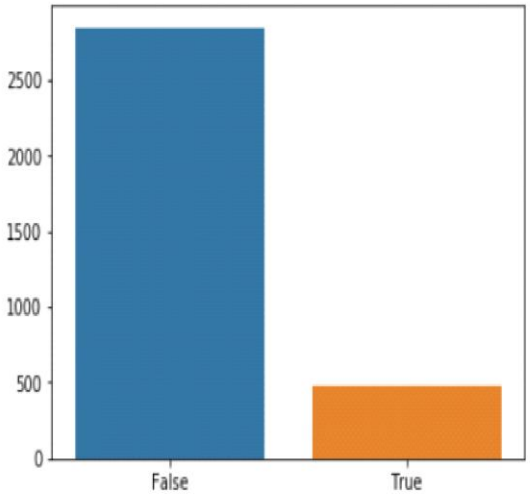

Fig. 2.Unbalanced Dataset Attributes

Figure 2 demonstrates how the dataset was unbalanced with respect to the given attributes that are used for predicting churn.

MODEL DESIGNING

Machine learning model is a mathematical representation. It can be represented to a real-world process. In the training data, this learning algorithm searches / finds patterns so that parameters of input will correspond to the given target. The predictions can be made using the output of the training process that results in a model of machine learning. Training a model simply means learning (determining) good values for all the labelled examples ie. having good values for weights and the bias from labelled examples. In the method of supervised learning, a machine learning algorithm will build a model by examining many examples. It attempts to find a model that will result in minimization of loss.This process is termed as empirical risk minimization.

IV. ALGORTHMS USED

SUPPORT VECTOR MACHINE

Here in this methodology needs to find a hyper plane and this algorithm will help in classification on data points. This hyper plane can exist in any dimensional space with any number of features. In order to find the maximum marginal plane this needs to classify data points with higher or better value. This value is called as confidence. This maximum marginal plane is useful in predicting the difference and similarities in data points.

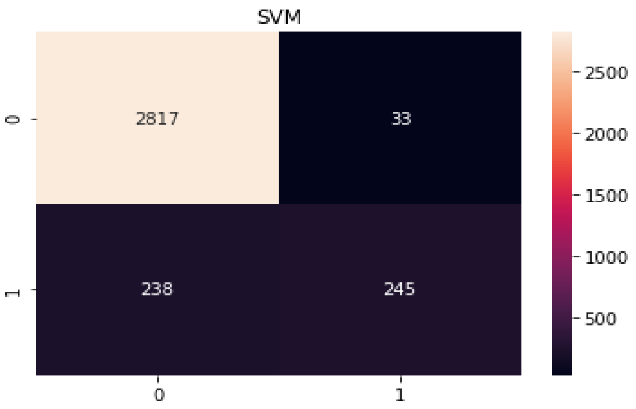

Fig. 3.Confusion Matrix for SVM

Figure 3 demonstrates the confusion matrix created based on the attributes in the dataset for support vector machine.

This can calculate or predict the hyper plane with the help of number of input features. The hyper plane can be a line or n dimensional plane. For eg. if the features count is two it will result in a line. But it will reach a plane like structure when the features count is two after that no one can't even imagine the nature of the hyper plane when count values is greater than two. Here comes the support vectors to find the inclination of the plane and its positional placement in the given space. This support vectors are very useful in finding the maximum marginal plane. SVM can be used with help of support vectors.

K-NEAREST NEIGHBORS

The KNN algorithm is very easy so that it can be implemented with less effort. It is a simple machine learning algorithm that belongs to the category of supervised algorithms. It can solve problems that fall under classification as well as regression. Here the things that are common rely on very near as well closer proximity. It can also say that common variables or things are near/ very close to each other. Here the common things can be like distance, nearness / closeness with basic maths like calculating distance between two pints in a graph.

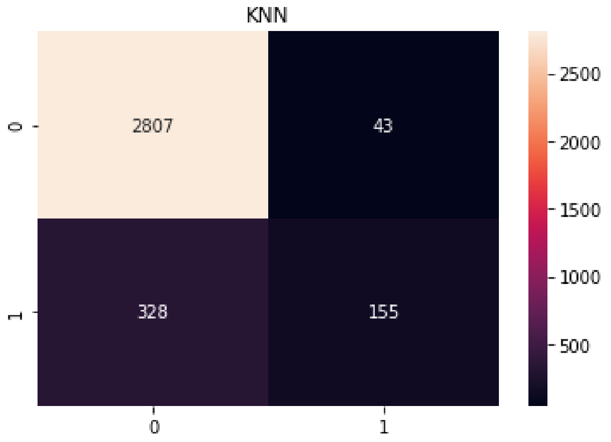

Fig. 4.Confusion Matrix for KNN

Figure 4 demonstrates the confusion matrix created based on the attributes in the dataset for KNN.

This method have to select the correct K value in order to make the predictions correct and accurate. This have to select the correct K value by using this algorithm. This correct K value will increase the prediction accuracy and reduce the incorrectness and errors. This algorithm is very simple so that it don't need any external assumptions or any other parameters .This property makes the algorithm better and a versatile one. This can be used for searching purposes but there may be difficulties when the data or variable increases.

LOGISTIC REGRESSION

Relapse examination is a type of prescient displaying method which explores the connection between a needy (target) and autonomous variable (indicator). This procedure is utilized for anticipating, time arrangement demonstrating and finding the causal impact connection between the factors. For instance, connection between rash driving and number of street mishaps by a driver is best concentrated through relapse. Strategic Regression is a Machine Learning grouping calculation that is utilized to foresee the likelihood of a clear cut ward variable. In strategic relapse, the reliant variable is a paired variable that contains information coded as 1 (truly, achievement, and so forth.) or 0 (no, disappointment, and so forth.). As it were, the strategic relapse model predicts P(Y=1) as a component of X. Calculated Regression is one of the most well known approaches to fit models for clear cut information, particularly for twofold reaction information in Data modeling.

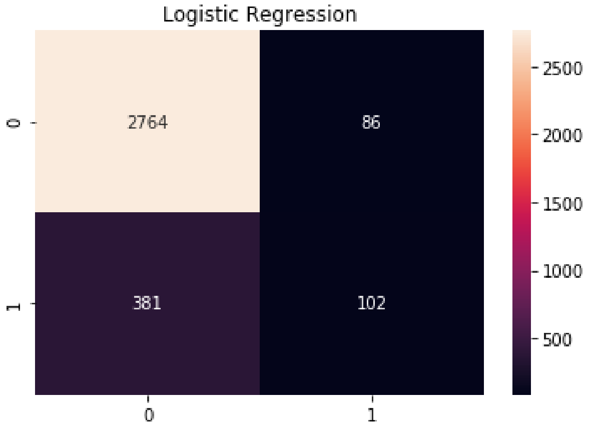

Fig. 5.Confusion Matrix for Logistic Regression

Figure 5 demonstrates the confusion matrix created based on the attributes in the dataset for Logistic Regression.

It is the most significant (and presumably generally utilized) individual from a class of models called summed up direct models. In contrast to straight relapse, calculated relapse can straightforwardly foresee probabilities (values that are confined to the (0,1) interim); besides, those probabilities are very much adjusted when contrasted with the probabilities anticipated by some different classifiers, for example, Naive Bayes. Calculated relapse protects the minor probabilities of the preparation information. The coefficients of the model likewise give some trace of the general significance of each info variable.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Strategic Regression is utilized when the reliant variable (target) is straight out. Strategic relapse is utilized in different fields, including AI, most clinical fields, and sociologies. For e.g., the Trauma and Injury Severity Score (TRISS), which is generally used to anticipate mortality in harmed patients, is created utilizing calculated relapse. Numerous other clinical scales used to survey seriousness of a patient have been created utilizing calculated relapse. Strategic relapse might be utilized to foresee the danger of building up a given sickness (for example diabetes; coronary illness), in light of watched attributes of the patient (age, sex, weight file, consequences of different blood tests, and so on.).

RANDOM FOREST

Random forest is one type of supervised learning algorithm that is used for regression as well as classification. Bu it is mainly for classification types of problems.A forest is made up of trees and more number of trees means that the forest is more robust. Similarly, here random forest algorithm creates decision trees on the samples of data and then gets prediction from each one of them and finally the best solution was selected by voting. It is one of the ensemble method that is better than a single decision tree because it will reduces over-fitting by result averaging.

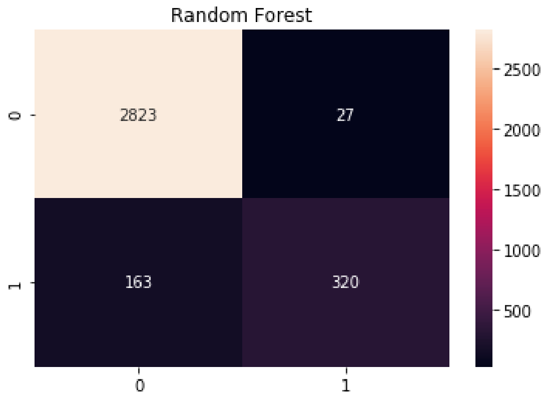

Fig. 6.Confusion Matrix for Random Forest

Figure 6 demonstrates the confusion matrix created based on the attributes in the dataset for Random Forest.

Here it can overcome over fitting problem by averaging or combining the results from various/different decision trees. For a large range of data items random forest works well rather than a single decision tree. The variance of random forest is very less compared to single decision tree. It gives high accuracy and it is very flexible. Random forest algorithm is useful for scaling of data. It has good accuracy even after providing data without any type of scaling. Random forest algorithm maintains very good accuracy even after missing large proportion of the data. The main disadvantage of Random forest algorithms are its complexity. Random forests construction was very hard and consumes more time than decision trees. Computational resources are required more for the implementation of random forest algorithm. When this have a collection of decision trees in large amount it is found that it is very less intuitive. The process of prediction in random forests consumes more time in comparison to other algorithms.

GRADIENT BOOSTING

The AI calculation Gradient Boosting is utilized for finding out about relapse and order issues, thus, the calculation delivers a forecast model as a powerless expectation models which is normally known as the choice trees. The calculation can be the most effectively clarified by first presenting the calculation called AdaBoost. The AdaBoost Algorithm can be characterized as,it starts via preparing a choice tree . Right now perception is allocated an equivalent weight. Angle Boosting strategy trains numerous models in a steady, added substance, and successive way.

Gradient boosting works better since it is a vigorous out-of-the-crate classifier (regressor) that can perform on a dataset on which negligible exertion has been spent on cleaning and can learn complex non-direct choice limits through boosting. In the event that you cautiously tune parameters, slope boosting can bring about great execution than irregular timberlands. In any case, inclination boosting may not be a superior decision in the event that you have a ton of clamor, as it can result in over fitting. This calculation likewise will in general be harder to tune than arbitrary timberlands.

AREA AND TOOL USED:

1. MACHINE LEARNING

Machine learning uses algorithms to analyse and predict the given data set and provide results so that the methodology can be modified in order to get desired output. Here this methodology uses machine learning to analyse the telecom dataset and provide the predicted churn value to improvise the performance of the telecom industries profit value. It mainly focuses whether the customer is satisfied or not.

2. GOOGLE COLAB

Google colab is a free cloud service available so that anyone can easily utilise and perform real world requirements in a very simple environment. This methodology use google colab to predict the data set that conatins nearly 3000 data and also 21 attributes. This methodology is done using google colab jupyter notebook that don’t requires any pre requisites to evaluate the dataset using any particular machine learning algorithm.

V. RESULT

Classifier	Precision	Recall	F1- score	Accuracy
SVM	0.91	0.92	0.91	92%
Logistic Regression	0.83	0.86	0.83	86%
KNN	0.88	0.89	0.87	89%
Random forest	0.94	0.94	0.94	94%
Gradient Boosting	**0.95**	**0.95**	**0.95**	**95%**

Fig.7. Summary of Results.

Figure 7 demonstrates the results generated using different algorithms for the given dataset from which the maximum accuracy from gradient boosting algorithm is chosen for accurate prediction of churn.

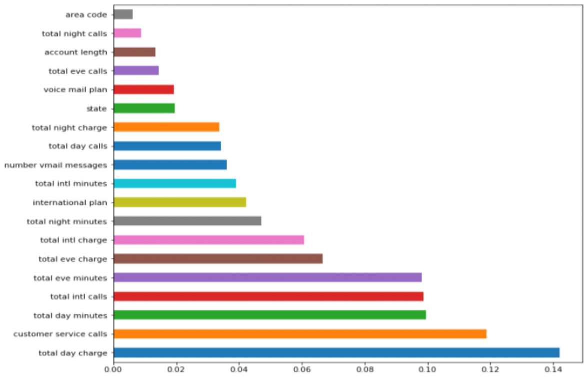

Fig.8. Feature Selection by Gradient Boosting

Figure 8 demonstrates the parameters in the dataset that contributes at different levels for resulting of churn in telecom.

VI. CONCLUSION

Data Analytics describes the procedure of retrieving pattern from large data set related to machine learning, data base, and statistics. Machine learning helps in deciding the line of treatment to be followed for the extraction of knowledge from various suitable databases. This type of research is useful in the telecom industry to help the companies to make more profit. It is known that churn prediction is one of the most important thing that results in one of the major reasons of income to telecom companies. The patterns found from data can be used to make better decisions, identify risks and opportunities for future. The result will be decreasing due to the non-stationary data model phenomenon. Hence this model needs training ie, each period of time. The use of the Social Network Analysis features enhances the results of predicting the churn in telecom.

REFERENCE

1. A. W. Ndung'U, "Modeling of churn behavior of bank customers using logistic regression".
2. S.-P. Chiu, C.-C. Yang, and W.-C. Chu, "Evaluating factors for customer churn of hairdressing industry based on modified delphi method," in 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). IEEE, 2018, pp. 1–2.
3. J. T. Wei, M. C. Lee, H. K. Chen, and H. H. Wu, "Customer relationship management in the hairdressing industry: An application of data mining techniques," Expert Systems with Applications, vol. 40, no. 18, pp. 7513–7518, 2013.
4. E. G. Castro and M. S. G. Tsuzuki, "Churn prediction in online games using players login records: A frequency analysis approach," IEEE Transactions on Computational Intelligence & Ai in Games, vol. 7, no. 3, pp. 255–265, 2015.
5. N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pp. 1659–1665, 2014. 0018-9545 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
6. A. Idris, A. Khan, and Y. S. Lee, "Genetic programming and adaboosting based churn prediction for telecom," in IEEE International Conference on Systems, Man, and Cybernetics, 2012, pp. 1328–1332.
7. W. Bi, M. Cai, M. Liu, and G. Li, "A big data clustering algorithm for mitigating the risk of customer churn," IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1270–1281, 2016.
8. Y. Zhang, S. He, S. Li, and J. Chen, "A novel framework for mitigating intra-operator customer churn in telecommunications,"
9. http://www.dbmarketing.com/telecom/churnreduction html.
10. I. Roos and A. Gustafsson, "The influence of active and passive customer behavior on switching in customer relationships,"

AUTHORS PROFILE

Ms. M. Kavitha Margret, presently working as an assistant professor in Sri Krishna College of Technology ,Coimbatore ,She had completed her UG Degree in Madurai Kamaraj University and PG Degree in Anna University Chennai .She is having Total Experience of 8 Years in Academic Institutions, She is Pursuing PhD., in Anna University,Chennai.

Ms. M. Monishapriyadharshini, IV year CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs120@skct.edu.in

Mr. S. Nathies, IV year CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs127@skct.edu.in

Mr. C. Sriram, IV year CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. Email: 16tucs230@skct.edu.in

Retrieval Number: E2931039520 /2020©BEIESP
DOI: 10.35940/ijitee.E2931.039520

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication