Advanced Churn Detection Model in Telecom Industry Using Machine Learning Techniques

P.TRIMURTHY ,Student, [M.Tech], Dept. of Computer Science & Engineering, Amrita sai institute of science and technology , A.P., India.

Dr.P.CHIRANJEEVI, Professor, Dept. of Computer Science & Engineering, Amrita sai institute of science and technology, A.P., India.

Abstract: As technology plays major role in recent days, business became highly increased day by day especially in the field of telecom domain, a large amount of data is being generated on daily basis due to increase number of customers and heavy usage from the customers. Many Telephone service companies often use customer churn or moving analysis and customer churn rates are one of their metrics in business churn detection because the cost of getting new customer is costlier than maintaining the existing ones. Decision analysers and executives need to know the explanations behind turnover or migration of the clients. This paper proposes customer model for churn customers to another service that utilizes some classification algorithms to see the moving of clients and gives the most affecting elements behind the turnover of clients in the telecom segment. Highlighting the most impacting features is performed by utilizing Recursive Feature Elimination(RFE) and Variance Inflation Factor(VIF).The proposed model classifies the churn clients information utilizing supervised classification algorithms, in which the Support Vector Machine(SVM) with Synthetic Minority Oversampling Technique(SMOTE) and Principal Component Analysis(PCA) i.e SVM+PCA+SMOTE calculation performed well with 97.11% exactness than logistic regression, logistic regression +PCA+SMOTE, Random Forest +PCA+SMOTE. This paper additionally recognized the most affecting elements for churning that are basic utilized in deciding the underlying factors of turnover. By knowing the huge churn factors from the information, the profitability of the organization increments. The proposed churn recognition model is assessed by utilizing metrics such as accuracy, sensitivity (true positive rates) and specificity (true negative rates). The outcome for our proposed churn detection model produced better churn classification using the SVM+PCA+SMOTE.

Keywords: Machine Learning, Support Vector Machine (SVM), Logistic Regression, Random Forest, Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), Variance Inflation Factor (VIF), Recursive Feature Elimination (RFE).

Introduction

Telecom industry went to be one of the quickest developing industry and has expanded drastically in the ongoing times. The goal of telecom company is to improve the productivity with their profits and stay alive in the market world. A customer lose happens when the clients are not satisfied with the services of any telecom company. It results in migration of customer by switching to other service providers. Due to these competitive rates of various service providers, customers often tend to switch between them. Churning impacts the overall reputation of a company which results in its company loss. As a result many attempts have been made in telecom industry to detect the churning customers before they actually leave a service provider. So churn recognition plays a vital role in the telecom sector as telecom operators have to maintain their customers and enhance their Customer Relationship Management (CRM). The most challenging job in CRM is to maintain the existing customers than getting new customers. So maintain the existing customers with the service needs of the customer. If there is dissatisfaction in the service of the company, there is a chance of churn. Our proposed model have the ability to detect the turnover customers and then find the reasons for their attrition to avoid loss of customers and provide the preventive cares for not to turnover from the current service provider. There are two types of payment in this telecom industry – post-paid (customers pay a monthly bill after using the services) and prepaid (customers pay with a certain amount in advance and then use the services). In the post-paid model, when customers want to turn over to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an chance of churn. Be that as it may, in the prepaid model, clients who need to change to another system can essentially quit utilizing the administrations with no notification, and

it is difficult to tell whether somebody has really moved or not agitated. Right now, consider the prepaid churn discovery model.

Customer behavior can be identified by using the following phases:

1) Good phase: In this phase, the customer is happy with the service and stays in that service.

2) Action phase: In this phase, the customer is unhappy with the service and stays in dilemma whether to churn or non-churn.

3) Churn phase: In this phase, the customer is said to be churned.

Due to advancements in the research of big data, there exist many data mining and machine learning techniques which can be used to analyze these types of telecom sector data. In this paper, we have used machine learning algorithms to detect the churners and non-churners. The term machine learning was first given by Arthur L Samuel[1]. Machine learning is subset of artificial intelligence which enables computers to learn from data without any intervention. Machine learning is used to construct the algorithms that can learn from data available and can be used to make predictions and detections on data. Here a model is built from the given input data which is then used to make predictions on new data. As the data being collected is drastically increasing each day, this calls for the need of machine learning.

A large volume of data is being generated in the telecom sector and the data contains duplicate values and missing values, which lead to the less result for detecting churners and non-churners. To handle these issues, data preprocessing methods are adapted to remove noise from data, which results high performance model detection. Perform Exploratory data analysis (EDA) i.e univariate and bivariate analysis on the features of the customer data. Splitting of data can takes place as 70% of training data and 30% of testing data. The feature selection is done using Recursive feature elimination (RFE) and Variance Inflation Factor (VIF) and detect the factors which are most affecting for churning of customers and build the model using logistic regression. Then build the model using SVM+PCA+SMOTE, logistic Regression +PCA+SMOTE, Random Forest +PCA+SMOTE. Evaluate the model using metrics such as accuracy, sensitivity and specificity.

Related work:

churn customers discovery has been performed utilizing different systems including AI, data mining, and hybrid strategies. These are utilized to distinguish the churn clients, recognize the most affecting highlights and improve the efficiency of the organization which helps in dynamic and CRM (customer relationship management). As expanding innovation numerous systems are utilized to examine the information and distinguish reasons for client churn. CRM can apply these methods to get their benefit in the sector[12].There is part of awkwardness happens subsequent to preprocessing of the dataset in this way, as of now the various systems of examining are applied to take care of the issue of the imbalanced datasets. The appropriateness of the examining methodologies to re-establish the balance between the classes of churn clients and non-churn client's arrangement has been performed. Specifically, the synthetic sampling algorithm called as the SMOTE (Synthetic Minority Oversampling Technique) [5] have been explored. The Support Vector Machine (SVM) is the supervised classification algorithm [3-11]. The SVM calculation is utilized for the different arrangement issues in various applications [10]. To develop the best SVM classifier it is necessary to find correctly the kernel function like linear kernel or poly kernel or RBF(radial basis kernel) and values of the kernel function parameters [10]. A model was developed by using logistic regression and Naïve Bayes classification algorithms to forecast the customer churn and it was shown that hybrid strategies gives the better performance than single application of classification algorithms [2]. So hybrid techniques like combination of different algorithms, sampling techniques and features reduction methods are used to get the better performance than individual classification algorithm.

Proposed Work:

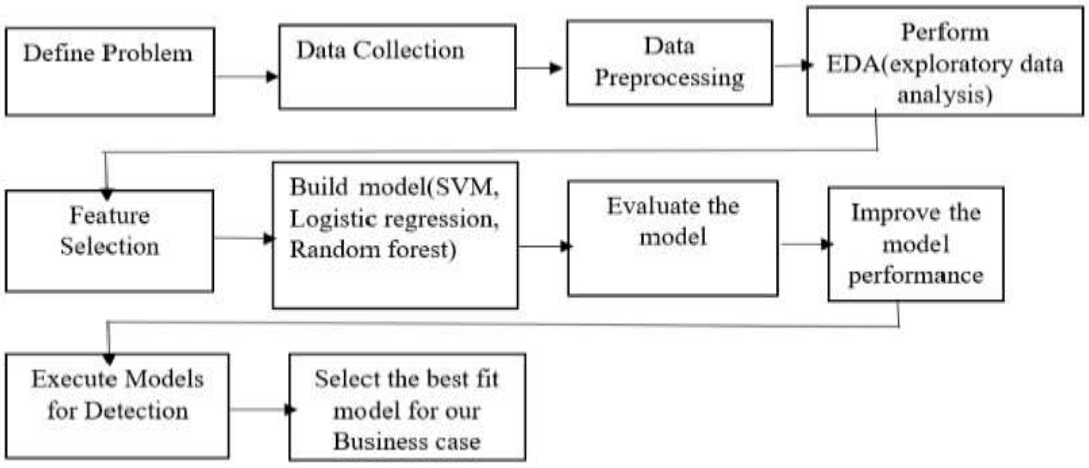

Figure 1: proposed customer churn detection model

This section represents the proposed churn detection model in telecommunication sector which are described in the steps as shown in above figure. The first step describes the problem statement, data collection and data preprocessing. In the next step EDA (Exploratory Data Analysis) takes place using univariate and bivariate analysis. In the next step feature selection can be done using RFE (recursive feature elimination) and VIF (variable inflation factor) in logistic regression model and dimensionality reduction (feature selection) can takes place using Principal component analysis (PCA) in SVM +SMOTE + PCA, logistic regression + SMOTE
+ PCA and Random Forest+SMOTE+PCA classification models. In next step we evaluate the models using accuracy, sensitivity and specificity. Compare the models that we built i.e logistic regression, logistic regression
+SMOTE+PCA, SVM+PCA+SMOTE and Random Forest +SMOTE+PCA models. Get the best churn detectionmodel for our telecom dataset.

3.1. Dataset description:

The dataset consists of null values and duplicate values with 99999 rows and 226 column features. But after thedata preprocessing, analysis and standardization of data we get 127column features and 30011 rows which thenfurther used to build the model. We us d the benchmark dataset which is available in online telecom churn file.

3.2. Data preprocessing:

It is very important for making the data without noisy data for the better performance of the model. There are a lot of missing values, duplicate values in the dataset and we can remove them using functions in python i.e dropping duplicate values and null values from the dataset. We analyzed the dataset using univariate and bivariate analysis and determines which attributes is having the strong relationship for customer churning by using correlation method. The data can be standardized if there is outliers present in the dataset.

3.3. Splitting the dataset:

We split the dataset as 70% of training data and 30% of testing data.

Dataset	Number of rows
Training data	21007
Testing data	9004
Total	30011

3.4. Features selection:

The most influencing features can be selected in this section using RFE function and VIF in logistic regression classifier. The most influencing features after performing RFE and VIF are as follows:

std_og_t2t_mou_8
loc_ic_mou_8
onnet_mou_8
loc_ic_t2m_mou_8
std_og_mou_8
total_ic_mou_7
total_rech_num_8
last_day_rch_amt_8
arpu_6
monthly_3g_8
spl_ic_mou_8
monthly_2g_8
sep_vbc_3g

Principal Component Analysis is the dimensionality or features reduction method which extracts the set of features from the large set of existing features in the dataset. Construct the Covariance matrix of the data set andcalculate cumulative covariance. The features with high variance can be selected as principal components. The dimensionalities can be reduced using cumulative covariance in SVM+PCA+SMOTE and Random Forest+SMOTE+PCA, logistic regression+SMOTE+PCA classification models.

3.5. Customer classification and detection:

There are two kinds of clients right now i.e churn clients and non-churn clients. In the proposed model we focused on churn clients and discover the reasons to migrate from one organization to the other organization. In the first step we evaluate the performance of logistic regression, logistic regression + PCA+SMOTE model using accuracy, sensitivity and specificity. In the next step we evaluate the performance of SVM+ PCA + SMOTE and Random forest +SMOTE+PCA models with 5-folds cross validation and determines the performance of the model using accuracy, sensitivity and specificity .At last we determine the best fit model for our business case i.e SVM+ PCA + SMOTE than other models. SVM with kernel function i.e RBF(Radial Basis Function) gives the best accuracy among SVM-poly and SVM-linear in which gamma value is 0.1 and cost is 10gets the best score with the highest accuracy of 97.11%.

Algorithms used in proposed churn detection model:

4.1. Logistic Regression:

The logistic regression model is used to classify the churn and non-churn customers. This is the best classification algorithm which maps the values between 0 and 1. In this we consider the threshold value as 0.5 based on the threshold we classify the data as churners and non-churners i.e if the value is more than 0.5 it is consider as 1(churn customers) otherwise 0(non-churn customers) . Logistic regression emphasis the sigmoid function.

4.2. SMOTE:

Synthetic minority oversampling technique (SMOTE) is a technique which is used for balancing the data of churners and non-churners. SMOTE is used to generate new minority class data from the existing data. The following steps are used to balancing of minority class instances:

1.let variable x be one of the minority class data2.find the

nearest member from x let say it as x`

3. calculate the value for new minority instance(y)=x + z(x-x`), where z is any random number between 0 and 1.

4. plot the new minority instance for balancing of data. Repeat the same process for every minority class instance.

4.3. SVM:

Supportvector machine (SVM) classifier is classification algorithm which is used to classify the churners and non-churners. The objective of the support vector machine algorithm is to find a hyperplane that distinctly classifies the churners and non-churners. We consider the hyperplane which is having the maximum margin distance from the support vectors. In this we used linear svm, poly svm and RBF (radial basis function) svm kernel.

4.4. Random Forest:

Random forest is an extension of bagging for decision makers that can be used for classification of churn customers and non-churn customers. This algorithm gives the most affecting features for churning the customers.

Evaluation of the model:

In this study, the proposed churn prediction model is evaluated using accuracy, sensitivity and specificity.accuracy: It identifies

several instances that were correctly classified.

Accuracy =(TP + TN)/(TP+TN+FP+FN)

where TN=True Negative(specificity)TP= True

Positive(sensitivity)

FN= False NegativeFP= False

Positive.

FP Rate gives that part of the data are incorrectly classified as positive rates in customer churn detection. FNgives part of data are incorrectly classified as negative rates in customer churn detection.

Sensitivity: It tells what portion of the data is correctly classified as positive. (true positive rate).

Sensitivity=(TP)/(TP+FN)

Specificity: It tells what portion of the data is correctly classified as negative. (true negative rate).

Specificity=(TN)/(TN+FP)

Receiver operating characteristic curve (ROC) area gives the average performance of all possible cost ratiosbetween False Positive rates and False Negative rates. If the ROC area value is equal to 1.0, this is a perfect prediction or detection which gives the highest accurate performance model .

Results:

The confusion matrix can is as follows:

predicted / actual	Non-churn customer(0)	Customer churn(1)
Non-churn customer(0)	True negative	False positive
Customer churn(1)	False negative	True positive

logistic regression:

	Non-churn customer(0)	Customer churn(1)
Non-churn customer(0)	15388	3796
Customer churn(1)	349	**1474**

accuracy=80.26%

sensitivity=80.85%

specificity=80.21%

logistic regression + PCA+SMOTE:

	Non-churn customer(0)	Customer churn(1)
Non-churn customer(0)	15874	3312
Customer churn(1)	2730	16469

accuracy=84.25%

sensitivity=85.78%

specificity=82.73%

SVM +PCA+SMOTE:

	Non-churn customer(0)	Customer churn(1)
Non-churn customer(0)	7917	315
Customer churn(1)	160	8059

Accuracy=97.11%

sensitivity=98.05%

specificity=96.17%

Random forest +SMOTE+PCA:

	Non-churn customer(0)	Customer churn(1)
Non-churn customer(0)	7047	1185
Customer churn(1)	1245	6974

Accuracy= 85.22%
sensitivity=84.85%

specificity=85.60%

From the above results we consider that best fit model for our dataset is SVM+PCA+SMOTE than any other algorithms that we used for our dataset. We get ROC curve area value approximately equal to 1.0 which can gives perfect detection.

Performance of classification Models:

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**
Logistic Regression	80.26	80.85	80.21
Logistic Regression +PCA+SMOTE	84.25	85.78	82.73
SVM+PCA+SMOTE	97.11	98.05	96.17
Random forest +PCA+SMOTE	85.22	84.85	85.60

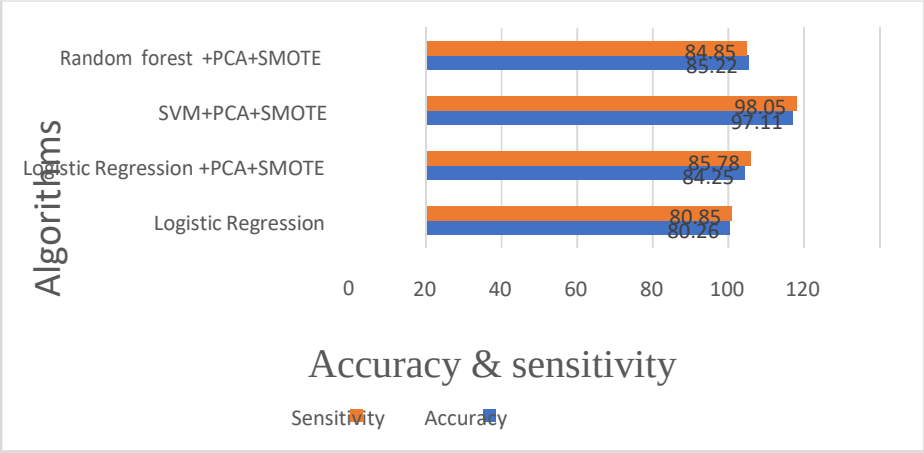

Figure 2: Accuracy and sensitivity performance in customer churn classification model

Conclusion:

In this paper we use the machine learning algorithms to classify the data as churn customers and non-churn customers. This also proposes a method to detect the churn customer for the dataset which decreases thetime complexity thereby improving the performance of the model. We did the comparative study on fourmodels thenwe get SVM+PCA+SMOTE algorithm performed well with 97.11% accuracy than logistic regression, logistic regression +PCA+SMOTE, Random Forest +PCA+SMOTE models. Clearly, Support Vector Classifier on datausing SMOTE oversampling technique and PCA for dimensionality reduction produced the most accurate detection.

References:

[1] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229. 1959.
[2] Rajnish Gupta (2019) Cadmium Nanoparticles And Its Toxicity. Journal of Critical Reviews, 6 (5), 1- 7. doi:10.22159/jcr.2019v6i5.34073
[3] Kenali, Norzaiti Mohd, Naimah Hasanah Mohd Fathil, Norbasyirah Bohari, Ahmad Faisal Ismail, and Roszaman Ramli. "Dental Development between Assisted Reproductive Therapy (Art) and Natural Conceived Children: A Comparative Pilot Study." Systematic Reviews in Pharmacy 11.1 (2020), 01- 06. Print. doi:10.5530/srp.2020.1.01
[4] Santulli, G.Epidemiology of cardiovascular disease in the 21st century: Updated updated numbers and updated facts(2013) Journal of Cardiovascular Disease Research, 1 (1), .
[5] N. Chawla, K. Bowyer, L.Hall, W. Kegelmeyer, JIAR, 16, 341-378 (2002)
[6] X. Gu, T. Ni, H. Wang, Scientific World Journal, 2014, 102-113 (2014)
[7] S. Zafeiriou, A. Tefas, I. Pitas, IEEE Transactions on Image Processing, 16(10), 2551-2564 (2007)
[8] R. Batuwita, V. Palade, IEEE Transactions on Fuzzy Systems, 18(3), 558-571 (2010)
[9] L. Demidova, I. Klyueva, A. Pylkin, 5-th Mediterranean Conference on Embedded Computing (MECO'2016), 322-325 (2016)
[10] L. Demidova, I. Klyueva, Y. Sokolova, N. Stepanov, N. Tyart, Procedia Computer Science, 103, 222- 230 (2017)
[11] L. Demidova, E. Nikulchev, Y. Sokolova, International Journal of Advanced Computer Science and Applications, 7(5), 294 (2017)
[12] C. Geppert, ''Customer churn management: Retaining high-margin customers with customer relationship management techniques,'' KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002