Machine Learning Approach on Customer Churn Analysis: A Comparative Study

Denisha. E[1], Shanmuga Sundari. B[2], Babu Renga Rajan. S[3]
[1, 2, 3] Dept of Computer Science & Engineering
[1, 2, 3] PET Engineering College, Tirunelveli,India.

Abstract- *The customer churn prediction is one of the challenging problems in the telecom industry. With the advancement in the field of machine learning and artificial intelligence, the possibilities to predict customer churn has increased significantly. We present a comparative study on the most popular machine learning methods applied to the challenging problem of customer churning prediction in the telecommunications industry. In the first phase of our experiments, all models were applied and evaluated using cross-validation on a popular, public domain dataset. In the second phase, the performance improvement offered by sampling techniques was studied. The data has been split into two parts train and test set in the ratio of 80% and 20% respectively. Finally, the overall analysis of confusion matrix of SVM, SVM with SMOTE-ENN and SVM with SMOTE Tomek are resulted 72.8%, 89.46% and 82.2% respectively. For the Support Vector Machines showed that the uses of SMOTE-ENN algorithm achieve highest accuracy. Continuation of this research will aim to improve performance of accuracy and other parameters compared to SVM models and selected Machine Learning Algorithms of Logistic Regression and Naive Bayes.*

Keywords- Telecommunication Industry, Customer churn, Customer Retention, Prediction Mode

I. INTRODUCTION

The rapid growth of the telecommunication sector makes it to have a very large customer base and makes it to operate in a highly competitive and quickly changing environment. In order to sustain in such an environment, the telecommunication sectors are taking more marketing efforts. However, these sectors are facing a number of issues due to their enormous growth [1]. In this perspective, the following are the most important research issues which need to be given more attention in telecommunication – Customer retention Churn Prediction Insolvency Prediction and Fraud Detection.

Today, all over the world the telecom sectors are facing huge revenue loss, due to the heavy competition and the churn behaviour of its customer's [2].Such behaviour of the customers makes an unwanted financial burden on the telecom sector, which in turn leads to huge revenue loss. By considering the competitive market as an advantage, many telecom sectors are investing more to acquire many customers in order to expand their business. Hence, telecom sectors are in a position to get back the investment and also in a need to gain a minimum profit within a short span of time. To sustain in the market, telecom sectors offer more schemes to attract the new customer. This is the main reason, which diverts the customers to leave from one service provider to another and such behaviour is called customer churn.In other words, Customer Churn is the term that refers to the customers who are switching from one service provider to another Growing number of such customers are becoming critical for the telecommunication sector, because the cost of attracting a new customer is higher than retaining the existing customer.

Data Mining - Telecom is an early adopter of the data mining techniques due to its large volume of data and its need to convert the same into useful information and knowledge. Data mining is a widely accepted interesting research area to mine the data in order to acquire the meaningful hidden knowledge from it [3]. Data mining not only performs the process of extracting the knowledge from data; it also performs the process of Data Cleaning, Data Integration, Data Transformation, Pattern Evaluation and Data Presentation.

Problem Statement - The business in telecom sector is intensively competitive among the service providers due to its rapid development. In such a market, companies are facing difficulties to get new customers. So, the approach is tuned from market development to product development by introducing new Value Added Services (VAS) to attract the customer. This in turn binds the customers with other competitors and that causes customer churn. Due to this, customer churn becomes the key concern in telecommunication. So, prediction of such customers in advance is very essential in order to retain them to avoid revenue loss. In addition to this, determining the factors that cause the customer churn and prioritizing the same help the telecom sector to prevent the customer churn [4]. Hence a model is needed which predicts the customer churn in telecom sector. The model should also have the ability to identify the factors which influence the churn.

Summary of our contribution - We have applied Over-sampling and Under-sampling algorithm to perform feature selection and to reduce the dimensions of the data-set. After, pre-processing of data, we have applied some of the famous machine learning techniques which are used for predictions like logistic regression, SVM, etc. and k-fold cross validation has been performed to prevent over fitting. Then we have use the power of ensemble learning in order to algorithms and achieve better results. Then we have evaluate the algorithms on test set using confusion matrix and AUC curve, which have been mentioned in form of graphs and tables in order to compare which algorithm performs best for this particular data-set.

II. REVIEW OF LITERATURE

Many approaches were applied to predict churn in telecom companies. Most of these approaches have used machine learning and data mining. The majority of related work focused on applying only one method of data mining to extract knowledge, and the others focused on comparing several strategies to predict churn.

Idris proposed an approach based on genetic programming with AdaBoost to model the churn problem in telecommunications. The model was tested on two standard data sets. One by Orange Telecom and the other by cell2cell, with 89% accuracy for the cell2cell dataset and 63% for the other one [5]. Huang et al. studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China's largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC [6]. Makhtar et al. proposed a model for churn prediction using rough set theory in telecom. As mentioned in this paper Rough Set classification algorithm outperformed the other algorithms like Linear Regression, Decision Tree, and Vot Perception Neural Network. Various researches studied the problem of unbalanced data sets where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction problem [7].

Amin et al. compared six different sampling techniques for oversampling regarding telecom churn prediction problem. The results showed that the algorithms (MTDF and rules-generation based on genetic algorithms) outperformed the other compared oversampling algorithm [8].

Burez and Van den Poel studied the problem of unbalance datasets in churn prediction models and compared performance of Random Sampling, Advanced Under-Sampling, Gradient Boosting Model, and Weighted Random Forests. They used (AUC, Lift) metrics to evaluate the model. The result showed that under sampling technique outperformed the other tested techniques [9].

Ahmed et al., author's applied tree based algorithms for the customer churn prediction, namely, decision tree, random forest, GBM tree algorithm, and XGBoost. In comparative analysis, XGBoost performed superior than others in terms of AUC accuracy. However, accuracy can be further improved using the optimization algorithms for the feature selection process [10].

Most of the previous research papers did not perform the feature engineering phase or build features from raw data while they relied on ready features provided either by telecom companies or published on the internet. In this paper, the feature engineering phase is taken into consideration to create our own features to be used in machine learning algorithms. We prepared the data using a big data and compared the results of three classification based machine learning algorithms.

III. PROPOSED SYSTEM

This consists of various phases of the proposed model. It consists of five phases, namely, **Phase 1:** Identification of most suitable data (variance analysis, correlation matrix, outlier removal, etc.), **Phase 2:** Cleaning & Filtering (handling null and missing values) and **Phase 3:** Feature Selection (using GSA). **Phase 4:** Development of predictive models (Logistic Regression, SVM, Naive Bayes, etc.). **Phase 5:** Cross validation (using k-fold cross validation). Finally, the evaluation of predictive models on test set (using Confusion matrix & AUC curve) has been presented in **phase 6 (Figure 3.1)**.

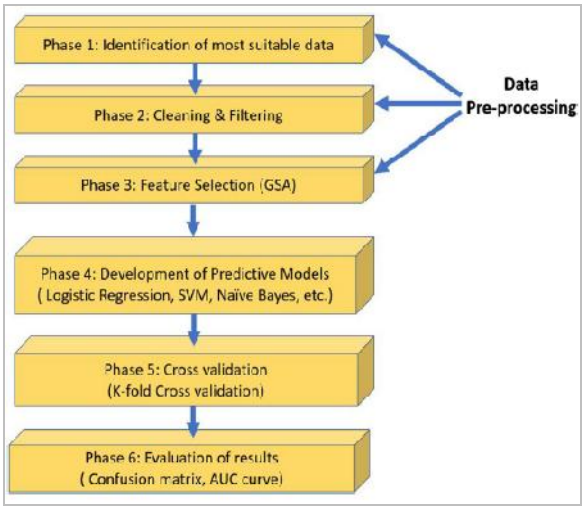


Fig. 3.1 Framework proposed system.

Data pre-processing is one of the important techniques of data mining which helps to clean and filter the data [11]. Thus, removing the inconsistencies and converting raw data into a meaningful information which can be managed efficiently. It is important to remove null values or missing values in the data-set and to check the data-set for imbalanced class distributions, which has been one of the emerging problems of data mining. The problem of imbalanced data-set can be solved through re-sampling techniques, by SMOTE, etc.

Phase 1: Identification of most suitable data: In order to establish a customer churn predictive model, firstly, select the important data or information from raw data in order to develop an efficient predictive model [12]. For identification of important data variance analysis has been adopted. Then correlation matrix is used to study the intra-relationship between the attributes. For class balancing dummy rows have been added by using re-sampling techniques.

Phase 2: Cleaning & Filtering: This phase consists of data cleaning and filtering by removing missing values, non-relevant parameters, etc. Data cleaning is the key to reduce dimensions of the data-set [13]. As the dimension increases, more time and power of computation is required. In the proposed methodology, data visualization is taken into consideration for understanding or extracting deeper insights from the data.

Phase 3: Feature Selection: In imbalance available pattern recognition, feature extraction plays an important role often in reducing the misclassification error and improving the classification performance. It's improve the performance of churn prediction in telecommunication industries with class-imbalanced data, two well-known oversampling techniques e.g. SMOTE-ENN and Tomek have been used to balance the imbalanced data [14].

Phase 4: In this phase predictive models are applied to make predictions. In order to optimize the results obtained from various classifiers, we have applied some existing techniques, namely, ensemble learning. Therefore, in the proposed methodology various models are applied, namely, Logistic Regression, Naive Bayes and SVM Classifier to make the predictions. Further the models and their respective hyper parameters have been fine-tuned using k--fold cross-validation.

Phase 5: It's a re-sampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called as k, which refers to the number of splitted groups in a given data sample. The k - Fold Cross Validation shuffles the data-set randomly, then splits the train set into k groups. From the splitted groups one group is randomly chosen as a test set and remaining as train sets.

Phase 6: Model evaluation is the key for analysing the performance of the proposed model. For model evaluation confusion matrix and AUC curve are taken into consideration. Then we have compared the results in order to identify the best performing model for the data-set.

IV. METHODOLOGY

Classification of modelling often encountered with an imbalanced dataset problem, where the number of majority class is much bigger than the minority class, thus make the model unable to learn from minority class well. This becomes a serious problem when the information in the dataset from the minority class is more important. One of the solutions to overcome that weak is to generate new examples that are synthesized from the existing minority class. This method is well known as Synthetic Minority Oversampling Technique or SMOTE. There are two variations of SMOTE implemented in this article, SMOTEENN and SMOTE-Tomek Links method.
It is a way the hidden features that are present in the rows and columns of data by visualizing, summarizing and interpreting of data. Some of the data visualizations can be seen in following Table 4.1 and Figures 4.1.

Table 4.1 **Different variables and data types**

Si. No	**variables**	**data types**
1	gender	object
2	Senior Citizen	int64
3	Partner	object

4	Dependents	object
5	Tenure	int64
6	Phone Service	object
7	MultipleLines	object
8	InternetService	object
9	OnlineSecurity	object
10	OnlineBackup	object
11	Device Protection	object
12	Tech Support	object
13	Streaming TV	object
14	Streaming Movies	object
15	Contract	object
16	Paperless Billing	object
17	Payment Method	object
18	Monthly Charges	float64
19	TotalCharges	float64
20	Churn	object

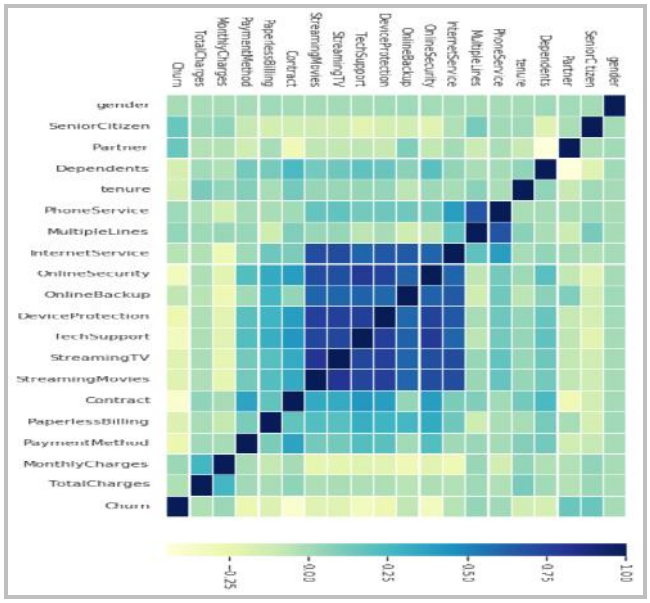


Fig. 4.1Correlation of all features in Pearson method.

Machine learning models - In the following, three well casted and popular techniques used for churn prediction has been presented succinctly, under the canopy of facts considered such as reliability, efficiency, and popularity in the research community.

Regression analysis (LR) is a probabilistic statistical classification model. It is also used for binary classification or binary prediction of a categorical value (e.g., house rate prediction, customer churn) which depends upon one or more parameters (e.g., house features, customer features). In addressing the complex problem of customer churn prediction problem, data first has to be casted under proper data transformation from the initial data in order to achieve good performance and sometimes it performs as good as Decision Trees [15].

Naive Bayesian classifiers assume that the value of each feature has an independent influence on a given class, and this assumption is called class conditional independence that is used to simplify the computation, and in this sense, we call it "Naive" [16]. In simple terms that this classifier assumes that the presence of feature vector (customer churn) is independent from the other feature vectors that are present in the class. The Naïve Bayes classifier is not regarded as a good classifier for large data-set but as our data-set was only about 7000 instances. It showcased good results.

Support Vector Networks introduced by Boser, Guyon, and Vapnik are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It divides the prediction into two parts +1 that is right side of the hyperplane and –1 that is left side of the hyperplane. The hyperplane is of width twice the length of margin. Depending on the type of data i.e. (scattered on the graph) tuning parameter like kernels are used like linear, poly, rbf, callable, pre-calculated [17].

Performance indicators - Recall - It is the ratio of real churners (i.e. True Positive), and is calculated under the following:

$$Recall = \frac{T_p}{T_p + F_n}$$

Precision - It is the ratio correct predicted churners, and is calculated under the following:

$$Precision = \frac{T_p}{T_p \; I \; F_p}$$

Accuracy - It is ration of number of all correct predictions, and is calculated under the following:

$$Accuracy = \frac{(T_p + T_n)}{(T_p + F_p + T_n + F_n)}$$

F –measure - It is the harmonic average of precision and recall, and it is calculated under the following:

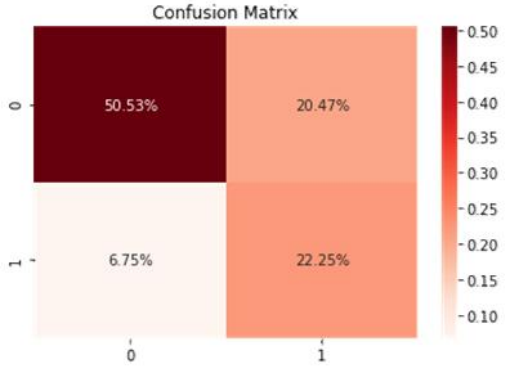
$$F - measure = \frac{(2\,X\,Presicion\,X\,Recall)}{(Precision + Recall)}$$

V. DISCUSSION

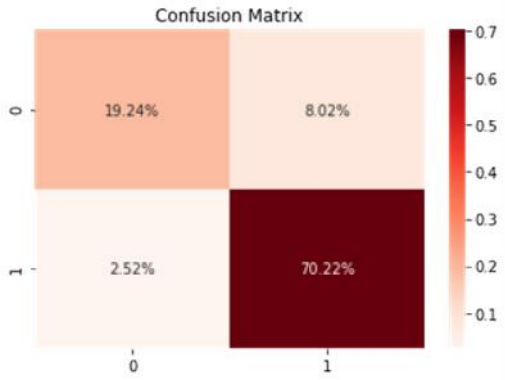
Final binary pre-processed dataset implemented on chosen standard Machine Learning algorithm of Support Vector Machines. Furthermore, accuracy performance of above model improved with help of SMOTE-ENN and SMOTE-Tomek algorithms. The acquired, evaluated results are represented in Table 5.1. These results are graphically represented confusion matrix shown in figure 5.1 (a), (b) and (c) respectively.

Table 5.1 Evaluation of Support Vector Machine models

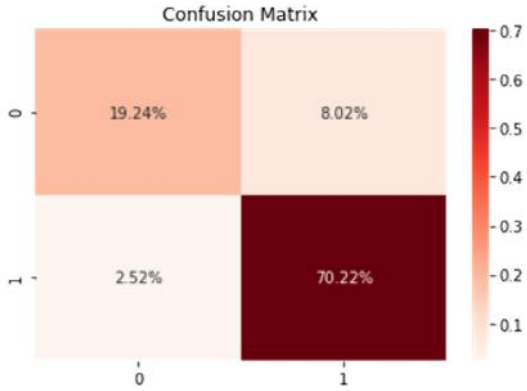
Model	Accuracy	Precision	Recall	F1-Measure
SVM	72.78	88.21	71.17	78.78
SVM (SMOTE-ENN)	89.46	88.42	70.5900	78.5
SVM (SMOTE-Tomeklink)	82.2	90.52	71.71	80.02



(a)



(b)



(c)

Fig 5.1 (a), (b) and (c) Graphical representation of Confusion matrix of SVM, SVM with SMOTE-ENN and SVM with SMOTE-Tomek.

\In overall analysis of confusion matrix of SVM, SVM with SMOTE-ENN and SVM with SMOTE Tomek are resulted 72.8%, 89.46% and 82.2% respectively. For the Support Vector Machines showed that the uses of SMOTE-ENN algorithm achieve highest accuracy. Especially, highest F1- measure were observed compared to others. Continuation of this research will aim to improve high accuracy of confusion matrix and other parameters compared to SVM models using of selected Machine Learning Algorithms of Logistic Regression and Naive Bayes.

VI. FUTURE FINDING AND CONCLUSION

Nowadays, as a result of extensive development and growth of telecommunication industry, it requires large companies who must be familiar with understanding the customers and their aspirants. One of the most important concerns of larger telecom industry is customer churn due to lagging of customer satisfaction and high competitive services. Therefore, companies are seeking to find customer churn influencing factors and its necessary actions to reduce the customer churn.Many researches have been developed to address solution for this issue. The main contribution of this project has following objectives: to identify and track the influencing behavioural factors of churn using customer's related dataset and to develop a customer churn predication model using machine learning standardized algorithm for obtaining best results. The accuracy of results will be verified and confirmed through confusion matrix. The analysis and investigation results will shows a great impact and performance in revenue potential for telcom industry.

VII. ACKNOWLEDGMENT

I would like to express my sincere thanks to Babu Renga Rajan.S for his valuable guidance and support in complete this article. I would also like to express my gratitude towards our College.

REFERENCES

[1] Debnath, R. M., & Shankar, R. (2008). Benchmarking telecommunication service in India: an application of data envelopment analysis. Benchmarking: An International Journal.

[2] Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., &Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-4). IEEE.

[3] Shaw, M. J., Subramaniam, C., Tan, G. W., &Welge, M. E. (2001). Knowledge management and data mining for marketing. Decision support systems, 31(1), 127-137.

[4] Richter, Y., Yom-Tov, E., &Slonim, N. (2010, April). Predicting customer churn in mobile networks through analysis of social groups. In Proceedings of the 2010 SIAM international conference on data mining (pp. 732-741). Society for Industrial and Applied Mathematics.

[5] Idris, A., Khan, A., & Lee, Y. S. (2012, October). Genetic programming and adaboosting based churn prediction for telecom. In 2012 IEEE international conference on Systems, Man, and Cybernetics (SMC) (pp. 1328-1332). IEEE.

[6] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., ...& Zeng, J. (2015, May). Telco churn prediction with big data. In Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 607-618).

[7] Makhtar, M., Nafis, S., Mohamed, M. A., Awang, M. K., Rahman, M. N. A., &Deris, M. M. (2017). Churn classification model for local telecommunication company based on rough set theory. Journal of Fundamental and Applied Sciences, 9(6S), 854-868.

[8] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ...&Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access, 4, 7940-7957.

[9] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[10] Ahmad, A. K., Jafar, A., &Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[11] Hariharakrishnan, J., Mohanavalli, S., & Kumar, K. S. (2017, January). Survey of pre-processing techniques for mining big data. In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP) (pp. 1-5). IEEE.

[12] Mishra, N., &Silakari, S. (2012). Predictive analytics: a survey, trends, applications, oppurtunities& challenges. International Journal of Computer Science and Information Technologies, 3(3), 4434-4438.

[13] Lalwani, P., Mishra, M. K., Chadha, J. S., &Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271-294.

[14] Jonathan, B., Putra, P. H., &Ruldeviyani, Y. (2020, July). Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek. In 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) (pp. 81-85). IEEE.

[15] Vafeiadis, Thanasis, Konstantinos I. Diamantaras, George Sarigiannidis, and K. ChChatzisavvas. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[16] Safitri, A. R., & Muslim, M. A. (2020). Improved accuracy of naive bayes classifier for determination of customer churn uses smote and genetic algorithms. Journal of Soft Computing Exploration, 1(1), 70-75.

[17] Xia, G. E., & Jin, W. D. (2008). Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice, 28(1), 71-77.