Grenze International Journal of Engineering and Technology, Jan Issue

Implementation of Chrun Rate Prediction System using Machine Learning

Veena Mohan M[1] and C K Vanamala[2]
[1]Student , Department of ISE, National Institute of Engineering, Mysuru, Karnataka, India.
[2]Associate Professor, Department of ISE, National Institute of Engineering, Mysuru, Karnataka, India.
Email: veenamohan48@gmail.com

***Abstract*— In the Telecommunication and entertainment Industry, customer churn detection is one of the most important research topics that the company must deal with retaining on-hand customers. Churn means the loss of customers due to existing offers from the competitors or maybe due to network issues. In these types of situations, the customer may tend to cancel the subscription to a service. The Churn rate has a substantial impact on the lifetime value of the customer because it affects the future revenue of the company and the length of service. Due to a direct effect on the income of the industry, the companies are looking for a model that can predict customer churn. The model developed in this work uses machine learning techniques. By using machine learning algorithms, we can predict the customers who are likely to cancel the subscription. Using this, we can offer them better services and reduce the churn rate. These models help telecom services to make them profitable. In this model, we used Logistic Regression, Random Forest, and SVM for churn rate prediction.**

***Index Terms*— Churn rate, Machine Learning, SVM, Random Forest, Logistic Regression, Subscription, customer churn.**

I. INTRODUCTION

The customer who ceases a product or service for a given period is referred to as churner. Customer churn is an imperative issue that is frequently connected with the existing cycle of the business [1]. At the point when the business is in a development period of its life cycle, deals are expanding exponentially and the number of new clients to a great extent dwarfs the number of churners. On the opposite side, organizations in a developed period of their life cycle set their emphasis on decreasing the rate of customer churn [2].
In this competitive era, it has become mandatory to maximize the profits periodically, for that various strategies have been proposed, namely, acquiring new customers, up selling the existing customers & increasing the retention period of existing customers. Among all the strategies, retention of existing customers is the least expensive compared to others. In order to adopt the third strategy, companies have to reduce the potential customer churn i.e., customer movement from one service provider to another. The main reason for churn is the dissatisfaction of the consumer service and support system. The key to unlocking solutions to this problem is by forecasting the customers which are at risk of churning [3]. One of the main aims of Customer Churn prediction is to help in establishing strategies for customer retention. Along with growing competition in markets for providing services, the risk of customer churn also increases exponentially. Therefore, establishing strategies to keep track of loyal customers (non-churners) has become a necessity. The customer churn models aim to identify early [4] churn

Grenze ID: 01.GIJET.10.1.382
© Grenze Scientific Society, 2024

signals and try to predict the customers that leave voluntarily. Thus, many companies have realized that their existing database is one of their most valuable assets [5] and according to researchers churn prediction is a useful tool to predict customers at risk.

Service companies suffer, particularly from the loss of valuable customers due to competitors known as customer churn. The scientific progress and the growing number of operators increased the level of opposition. Companies are working hard to survive in this aggressive market, depending on complicated strategies. The customer churn causes a considerable loss of Industry services and becomes a severe problem.

II. LITERATURE REVIEW

Horia Beleiu et al. [1], they adopted three machine learning approaches, namely, neural network, support vector machine and bayesian networks for customer churn prediction. In the feature selection process, principal component analysis (PCA) is taken into consideration to reduce the dimensions of the data. But, the feature selection process can be improved using optimization algorithms which increases the classification accuracy. In the performance evaluation, gain measure and ROC curve were used.

J. Burez et al. [2], authors tried to capture the class imbalance problem. They applied logistic regression and random forest with re-sampling technique. In addition, boosting algorithms were also applied. In the performance analysis, AUC and Lift are taken into consideration. They also observed the effect of advanced sampling techniques such as CUBE, but the obtained outcome did not improve the performance. However, the class imbalance problem can still be solved in a better way by using the optimization based sampling techniques.

Irfan Ullah et al., [3] identified churn factors that are essential in determining the root causes of churn. By knowing the significant churn factors from customers' data, Customer Relationship Management (CRM) can improve productivity, recommend relevant promotions to the group of likely churn customers based on similar behavior patterns, and excessively improve marketing.

K. Dahiya et al. [4], researchers applied the two machine learning models, namely, decision tree and logistic regression on churn prediction dataset. In experimentation, the WEKA tool was used. However, the problem can be solved in an efficient way by adopting other machine learning techniques.

Umman et al. [5], authors analyzed the mass database using logistic regression and decision tree machine learning models, but obtained accuracy was low. Therefore, further improvement is required for that other machine learning and feature selection techniques can be adopted.

Kavitha V et al., [6] used a Decision Tree, Random Forest, and XGBoost to predict the customers who are likely to cancel the subscription which can offer them better services and reduce the churn rate.By preprocessing and feature selection ,the data set for training and testing. For the above-mentioned algorithm, it is necessary to do some feature engineering to have more efficient and accurate results.

J. Hadden et al. [7], review of all the machine learning models taken into consideration as well as they presented deep analysis of existing feature selection techniques. In the prediction models, they found that the decision tree performed superiorly than others. In feature selection, optimization techniques also play a vital role that improves the prediction techniques. After the comparative analysis of existing techniques, authors suggested the path for the future research directions.

Y. Huang et al. [8], author's applied various classifiers on churn prediction dataset, in which the obtained results confirmed that random forest performs superior to others in terms of AUC and PR-AUC analysis. But accuracy can be further improved using the optimization techniques for the feature extraction

Krishna Sai and Sasikala [9] implemented an EDA using Visualization, statistical tests for feature selection and Data mining methods for predicting the likely churners by utilizing a Logistic Regression Model. Here dataset has been analyzed by using the data visualization techniques before entering the modeling process.

Idris [10] 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.

Aziz R. [11] proposed a model for churn prediction using rough set theory in telecom. As mentioned in this paper, the Rough Set classification algorithm outperformed the other algorithms like Linear Regression, Decision Tree, and Voted Perceptron Neural Network.

Burez and Van den Poel [12] studied the problem of unbalanced 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 the under-sampling technique outperformed the other tested techniques.

III. Description Of The Research

The system mainly focuses on giving the details about the customer churn, i.e., the loss of customers, which is one of the most common problems faced by enterprises or organizations. We are trying to propose a system which will predict customer churn. Further this system will help the scholars and business owners to be able to predict customer turnover. Using machine learning algorithms, we try to provide the accuracy of the results and make a comparative study on the SVM, random forest and logistic regression.

IV. Methodology

This section is the most important of the research article, as it gives the complete description of the implementation. Firstly, let us understand the dataset used in our prediction:

a. Dataset

As we know, the data set is the starting point for everything; it should have full-fledged data to make the machine learn about the problem. Datasets can be generated or developed from the scraped information available on the internet. Some issues we must create a dataset that makes sense that tells how to respond based on real-time inputs for the problem datasets can be gathered from the internet every day. Mock-up dataset based on trends found in real world case studies; 27,000 instances and 30 features. The dataset contains attributes related to the financial habits of the users. 70% of the data has been used for training and 30% of the data for testing and test data split in training data class 0 - 11075, class 1 - 7822. Class-imbalance is a problem so we will balance the classes count using resampling technique.

b. System Architecture

This section gives a detailed view of the process of churn rate prediction. Fig. 1 illustrates the same.

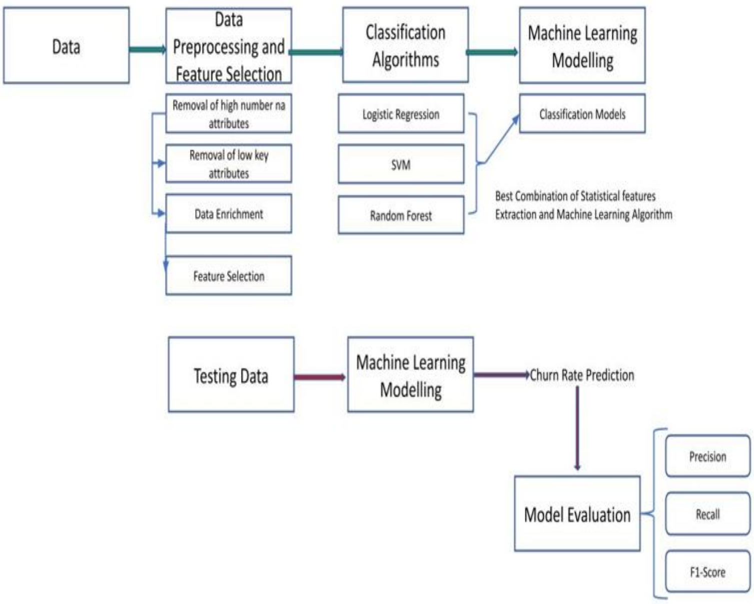

Fig. 1: Architecture for churn rate prediction

c. Data Processing

Data set is a collection of feathers and N number of rows. Many values are in different formats. In a dataset, there may be duplicate values or null values that may lead to some loss inaccuracy, and they may be dependent [13-15]. In Pre-processing non applicable attributes were removed because their score rate was very high.

d. Data filtering and Noise Removal

It is very crucial to make the data useful because unwanted or null values can cause unsatisfactory results or may lead to producing less accurate results [16]. In the data set, there are a lot of incorrect values and missing values. We analyzed the whole data set and listed out only the useful features. The listing of features can result in better accuracy and contains only valuable features. Low key attributes were also removed.

e. Feature Selection and Engineering

Feature selection is a crucial step for selecting the required elements from the data set based on the knowledge. The dataset used here consists of many features out of which we chose the needed features, which enable us to improve performance measurement and are useful for decision-making purposes while remaining will have less importance. The performance of classification increases if the dataset has only valuable variables and which are highly predictable. Thus, having only significant features and reducing the number of irrelevant attributes increases the performance of classification [17].

f. Prediction and Classification

Many techniques have been proposed for customer churn prediction in the telecommunication industry. These three modeling techniques are used as predictors for the churn prediction [18].

V. MODELS

The following are the models which are used for customer prediction:

a. *Forest:* It is one of the ensembles learning technique [12] which consists of several decision tree rather than a single decision tree for classification. While classifying all the trees in the random forest gives a class to an unknown example and the class having maximum votes will be assigned to the unknown example.

b. *SVM***:** In machine learning, Support Vector machine also Known as 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. A Support Vector Machine is a kind of classification technique, where the data points are separated by a line in case of linear SVM, and a hyperplane in case of non-linear SVM. The separation is chosen in such a way that the two sides of the hyperplane categorize the data set into two classes. When an unknown data comes it predicts which side/class it belongs to. The margin between the hyperplane and the support vectors are as large as possible to reduce the error in classification.

c. *Logistic Regression***:** It works on the divide and conquer approach. It is based on the random subspace method [19]. In this method several trees are formed, and each decision tree is trained by selecting any random sample of attributes from the predictor attributes set. By using logistic regression, we can predict the probability of a churn i.e., the likelihood of a customer canceling the subscription. Logistic regression is a supervised learning algorithm used for classification. In Logistic regression, we set a threshold; based on the limit, and only the classification is made using logistic regression. The threshold value is variable, and it is dependent on the classification problem.

d. *Performance indicators*

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

$$Recall = \frac{T_p}{T_p + F_n} \tag{1}$$

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

$$Precision = \frac{T_p}{T_p + F_p} \tag{2}$$

Accuracy**:** It is ratio of number of all correct predictions, and is calculated under the following:

$$Accuracy = \frac{(T_p + T_n)}{(T_p + F_p + Tn + F_n)} \tag{3}$$

VI. RESULTS AND DISCUSSIONS

We performed several experiments on the proposed churn model using machine learning algorithms on the dataset. In Fig.2, we can observe the results obtained while performing the experiment using the Random Forest algorithm and can check the accuracy. Random Forest (RF) is a useful algorithm that is suited for classification and can handle nonlinear data very efficiently [20]. RF produced better results and better accuracy and performance compared to the other techniques. As we need better accuracy to predict the customer churn, we prefer to use the technique, which results in better accuracy. Similarly, we can observe the results obtained when using the Logistic regression technique and SVM.

```
Accuracy with Logistic Regression Model:
0.6232868255340166
Logistic Regression: Confusion Matrix:
[[2562 2185]
 [ 866 2486]]
              precision    recall  f1-score   support

           0       0.75      0.54      0.63      4747
           1       0.53      0.74      0.62      3352

    accuracy                           0.62      8099
   macro avg       0.64      0.64      0.62      8099
weighted avg       0.66      0.62      0.62      8099
```

Fig.2: Accuracy of Logistic Regression model

```
Accuracy with RandomForest Classifier Model:
0.7144091863192987
RandomForest Classifier: Confusion Matrix:
[[3950  797]
 [1516 1836]]
              precision    recall  f1-score   support

           0       0.72      0.83      0.77      4747
           1       0.70      0.55      0.61      3352

    accuracy                           0.71      8099
   macro avg       0.71      0.69      0.69      8099
weighted avg       0.71      0.71      0.71      8099
```

Fig.3: Accuracy with Random Forest model

```
accuracy_score(y_test,pred)

0.6695888381281639

print(classification_report(y_test,pred))

              precision    recall  f1-score   support

           0       0.70      0.77      0.73      4747
           1       0.62      0.53      0.57      3352

    accuracy                           0.67      8099
   macro avg       0.66      0.65      0.65      8099
weighted avg       0.66      0.67      0.66      8099

cm=confusion_matrix(y_test,pred)
cm

array([[3660, 1087],
       [1589, 1763]])
```

Fig.4: Accuracy with Support Vector Machine

The precision and recall calculated with all the three algorithms are taken from the above outputs and are plotted in the below graph. Fig.5, illustrates the comparison of precision and recall values calculated with the Logistic Regression, Random Forest and Support Vector Machine algorithms.

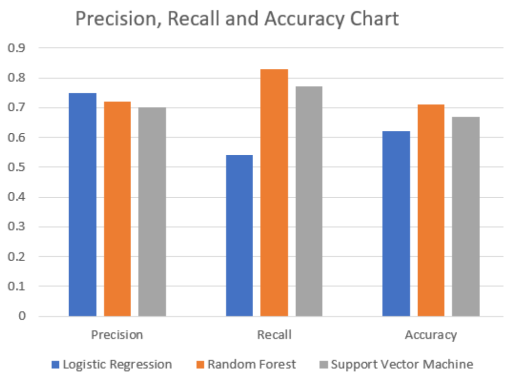

Fig.5: Comparsion Graph Precision, Recall and Accuracy of three Algorithms.

From the above comparisons we can derive the following, the Logistic Regression gives the best precision where as Random Forest works good with Recall as well as Accuracy.
Evaluate the Random Forest model based on their performance metrics like accuracy, precision and recall that is more important to identify churners than the non-churners accurately. Figure 3 shows the results obtained while performing the experiment using the Random Forest algorithm and can check the accuracy. This clearly shows that Random Forest performs better with non-linear data than other machine learning models.

TABLE I. MODEL AND THEIR ACCURACY

Model	**Accuracy**
Logistic Regression	62%
Random Forest	71%
Support Vector Machine	67%

VII. CONCLUSION

This paper provides a detailed study of the methods used for the process of customer churn prediction. The accuracy attained by several models is displayed in Table 6.4. Random Forest's classifier has a 62% accuracy rate. SVM achieved an accuracy of 66%, while another classifier we created, logistic regression, achieved an accuracy of 71 %. More accurate than the random forest classifier is logistic regression. Since the nature of the data makes it difficult to attain 100% prediction accuracy, we should always aim for it when building classifiers. The better your classifier works, however, the greater the prediction accuracy score.
A. Future research will consider enhancing the system performance using more sophisticated machine learning algorithms such a neural network, SVM, along with sophisticated assembly techniques like boosting, bagging.

REFERENCES

[1] Lalwani, Praveen & Mishra, Manas & Chadha, Jasroop & Sethi, Pratyush. (2022). Customer churn prediction system: a machine learning approach. Computing. 104. 1-24. 10.1007/s00607-021-00908-y.
[2] Burez J, Van den Poel D (2009) Handling class imbalance in customer churn prediction. Expert Systems with Applications 36(3):4626–4636
[3] Irfan Ullah , Basit Raza, Ahmad Kamran Malik , Muhamad Imran , Saif Ul Islam and Sung Won Kim., “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector”, In the proceedings of IEEE Access, vol. 07, no. 2169-3536, pp. 60134 - 60149, 2019.
[4] Dahiya, K., Bhatia, S.: Customer churn analysis in telecom industry. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 1–6 (2015).
[5] Umman et al. (2010) Customer churn analysis in telecommunication sector. ˙Istanbul Üniversitesi ˙I¸sletme Fakültesi Dergisi 39(1):35–49.
[6] Kavitha V, Hemanth G , Mohan S.V and Harish M , “Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms”, International Journal of Engineering Research & Technology(2278-0181), Vol. 9, Issue 05, pp. 181-184, 2020.
[7] Hadden J, Tiwari A, Roy R, Ruta D (2007) Computer assisted customer churn management: State-ofthe-art and future trends. Computers & Operations Research 34(10):2902–2917

[8] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J.: Telco churn prediction with big data. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp. 607–618 (2015)

[9] Krishna B.N, and Sasikala ,“Predictive Analysis and Modeling of Customer Churn in Telecom using Machine Learning Technique ,”In the proceedings of International Conference on Trends in Electronics and Informatics , Tirunelveli, India, pp. 6-11, 2019.

[10] Idris, A., Khan, A., Lee, Y.S.: Genetic programming and adaboosting based churn prediction for telecom. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1328–1332. IEEE (2012)

[11] Aziz R, Verma C, Srivastava N (2018) Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction. Annals of Data Science 5(4):615–635

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

[13] Asthana P (2018) A comparison of machine learning techniques for customer churn prediction. International Journal of Pure and Applied Mathematics 119(10):1149–1169

[14] Coussement K, De Bock KW (2013) Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research 66(9):1629–1636

[15] O. Ali and N. Z. Munauwarah, Factors affecting employee turnover in organization/Nur Zuhan Munauwarah Omar Ali, 2017.

[16] Syam, N.; Sharma, A. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Ind. Mark. Manag. 2018, 69, 135–146.

[17] Rombaut, E.; Guerry, M.A. Predicting voluntary turnover through Human Resources database analysis. Manag. Res. Rev. 2018, 41, 96–112.

[18] T.Vafeiadis, K.I. Diamantaras, G.Sarigiannidis, K.Chatzisavvas “Customer churn prediction in telecommunications”, Simulation Modelling: Practice and Theory 55 (2015) 1-9.

[19] T. Jahromi, M. Moeini, I. Akbari, and A. Akbarzadeh, ‘‘A dual -step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers,’’ J. Innov. Sustainab., vol. 1, no. 2, pp. 2179–3565.