An Efficient Technique for Feature Selection to Predict Customer Churn in telecom industry

Mr. Varun E
Assistant Professor
Department of CS &E
AIT, Chikmagalur
varune.ait@gmail.com

Dr. Pushpa Ravikumar
Professor & Head
Department of CS &E
AIT, Chikmagalur
flowersunpr@yahoo.co.in

Chandana S
CS&E Department
AIT College, Chikmagalur
chandana.s0908@gmail.com

Spandana K M
CS&E Department
AIT College, Chikmagalur
spandanakm26@gmail.com

Abstract

The evolution of technology has a great impact on the telecom industry, which has grown rapidly from telegraph to present high speed network. This rapid growth has resulted in the establishment of many telecom sectors which in turn has given rise to a stiff competition among them. Telecom sectors with improved technology needs to handle the large set of subscribed customer base. Now a days, in addition to acquisition of new customers to increase the company revenue, retaining the old customers is also found to be of much importance. So, all the telecom industries are concentrating on building a best predictive model in order to determine the churn rate. In this paper we mainly concentrate on refining the telecom dataset by applying the Pre-processing, feature selection and feature extraction techniques. The refined dataset is created to provide the prediction accuracy similar to or greater than the original dataset with less computation.

Keyword: *Feature selection, Feature extraction, Dimensionality reduction, Telecom, Prediction*

I. Introduction

Communication is an important part of everyone’s life whether it is long distance communication or short distance. It is required in every field. Communication media has evolved drastically from telegraph to the present high speed network. It has made everyone’s life simple and easy. The development of technology has given rise to the establishment of multiple telecom sectors. With the increase in the count of telecom industry the competition among them is also increasing rapidly. In addition to acquiring the new customers retaining the old customers has also become an important goal of many telecom companies. All the telecom sectors are after building the best predictive model to determine the customer who are about to churn. Churn refers to switching from one subscribed company to another. If the churn rate increases the revenue of the company decreases as it affects the new customers also. So it is very much important to monitor the old customers and identify the ones who are about to churn. If the identification is successful necessary measures can be taken to prevent them from subscribing to another company.

As the subscription takes place every day the details of the customer is huge to store. Determining the customer who is about to churn in this huge data requires lot of time. In this work we introduce the pre-processing techniques and the dimensionality reduction methods to reduce the size of the dataset which is to be trained to build the predictive model. In the first step pre-processing is done which involvesremoval of duplicate instances, handling missing values and elimination of irrelevant records. In the second step the processed data is taken to reduce the attribute count by identifying and eliminating the irrelevant attribute. It is done using the feature selection and feature extraction method.

II. Literature Survey

A.Feature selection:

Various feature selection and classification techniques can be used to determine the best attribute for prediction which is explained in [1]. It gives a detailed description about the tested dataset, software tool and the classification methods that are used. The paper [2] shows that by performing the data reduction and feature selection the prediction accuracy can be increased. It gives the ideology about the two important data preprocessing steps for churn prediction. The paper [3]gives the three deep neural network architectures and builds the corresponding churn prediction model using two telecom dataset. It concludes that the deep learning based models gives

978-1-7281-3241-9/19/$31.00 © 2019 IEEE

the better performance compared to the traditional classification models without using hand-picked features.In paper [4] the author has focused on different attribute selection methods for identifying the subset of attributes for churn prediction.

B. Churn prediction:

The paper [6] focuses on selecting the relevant data items which are really contributes to the specific analysis. It also focuses on the churn prediction, the importance of feature extraction and the use of data mining techniques in churn prediction using telecom dataset.

III. Methodology

The telecom dataset is collected from the Kaggle which consists of 21 attributes and 3430 instances. Description of each attribute is as follows:

- State: State attribute contains the 50 states and the district of Columbia
- Account length: This attribute consists of the integer value that represents from how long the customer account has been in an active state
- Area code: This Attribute represents the categorical variable for area code
- Phone number: A unique key for customer identification
- International Plan: This attribute has yes or no value that represents whether the customer has activated the international plan or not
- Voice Mail Plan: The yes or no value in this attribute gives an idea of voice mail plan activation
- Number of voice mail messages: Integer valued variablethat represents the number of voice mail messages sent
- Total day minutes: Continuous variable for number of minutes that the customer has used the service during day
- Total day calls: Represents the total number of day calls
- Total day charge: Represents the charges applied based on the total day minutes and total day calls
- Total evening minutes: Continuous variable value for number of minutes that the customer has used the service during evening
- Total evening calls: Represents the total number of evening calls
- Total evening charge: Represents the total charges applied based on the total evening minutes and total evening calls
- Total night minutes: Continuous variable value for number of minutes that the customer has used the service during night
- Total night calls: Represents the total number of night calls
- Total night charge: Represents the charges applied based on the total night minutes and total night calls
- Total international minutes: Continuous variable value for number of minutes that the customer has used the international service
- Total international calls: Represents the total number of international calls
- Total international charge: Represents the charges applied based on the total international minutes and total international calls
- Number of calls to customer service: This attribute represents the total number of calls made by the customer to the customer service
- Churn: class variable churn represented by booleanvalue that gives an idea to know whether the customer has churned or not based on the past data

The collected data is then reduced in terms of size using two steps:

1. Pre-processing
2. Dimensionality reduction

The figure 1 shows the detailed processing steps considered in this work.

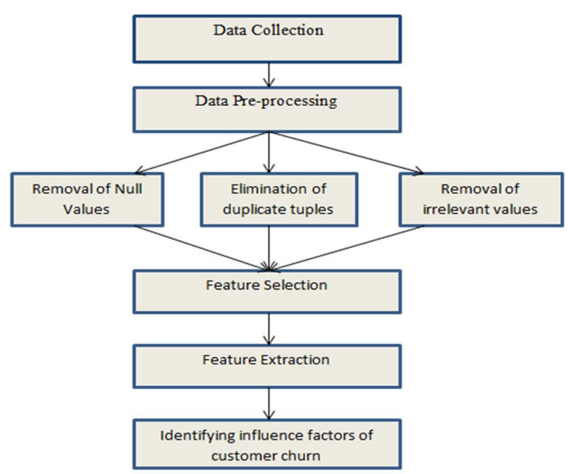

Figure 1: The Process Diagram

978-1-7281-3241-9/19/$31.00 © 2019 IEEE

A. Pre-Processing

Pre-processing refers to upgrading the available data in order to reduce the error rate and complexity.

It involves:

1. Removal of duplicate instances.

2. Handling of missing values

3. Eliminating irrelevant data

In this module the pre-processing is done which can perform all the above mentioned tasks. In removal of duplicate instances the redundant tuples are successfully identified in order to reduce the error rate. In second method the cells with null values are either ignored or replaced with a most appropriate values based on the analysis made on the existing values. In the third method the inconsistent values are removed.The outcome is a processed data which is fed as input to the dimensionality reduction module.

B. Dimensionality reduction

As the word suggests the dimensionality reduction refers to reducing the dimensions which is nothing but the attributes without affecting final outcome. The reduced data must provide the prediction whose accuracy is either similar or greater than that of the original dataset. The dimensionality reduction is done using two steps namely,

1. Feature selection
2. Feature extraction

In feature selection the attributes that are dependent on one another is identified using the correlation technique. This provides the set of attributes that are highly dependent on one another.

The Algorithm for feature selection is as follows:

Algorithm

Input: A set of numerical attributes 'A' where A = $\{A_1, A_2, A_3, \ldots, A_n\}$
Output: A set of linearly correlated attributes 'S'
Process:
For set of attributes in A
Begin
Form a subset each containing two attribute: a = $\{\{A_1,A_1\},\{A_1,A_2\},\ldots,\{A_i,A_j\},\ldots,\{A_n,A_n\}\}$
For each subset in 'a'
Begin
Evaluate $\sum A_i^2, \sum A_j^2$ and $\sum(A_i * A_j)$
Apply Karl Pearson Correlation Formula:

$$r = \sqrt{\frac{\sum(A_i * A_j)}{\sum A_i^2 * \sum A_j^2}}$$

if r = = +1
then there is a high positive correlation
else if r = = 0
then there is no correlation
else if r = = -1
then there is a high negative correlation
End
End

In feature extraction the dependent attributes are analyzed to eliminate the one that provides the same result as that of the attributes on which it is dependent which in turn reduces the size of the data.

IV. Results and discussion

A. Pre-Processing

During this method three steps are carried out. In the first step comparison of data is done in order to identify the redundant data. If any replication is found then it is immediately noted down to eliminate it. A graph is plotted to analyze the scattering of duplicate data in the dataset as shown in figure 2. In the second step each instance is searched for missing value. If it is found then the methods of replacement is executed to replace the missing value with most appropriate value. If the replacement is not possible then such instances are discarded. The graph plotted for missing values is as shown in figure 3. The last step is to eliminate the irrelevant value and it is done by determining the type of each value. If the type of the data is found to be irrelevant than such data are eliminated. The graph plotted for irrelevant data is as shown in figure 4.

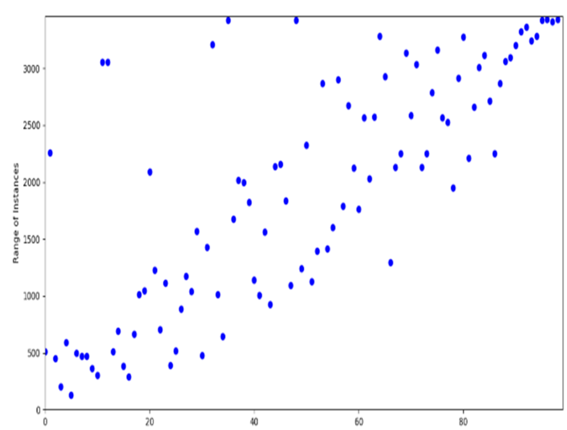

Figure 2: Scattered Graph plotted for determining duplicate instances

978-1-7281-3241-9/19/$31.00 © 2019 IEEE

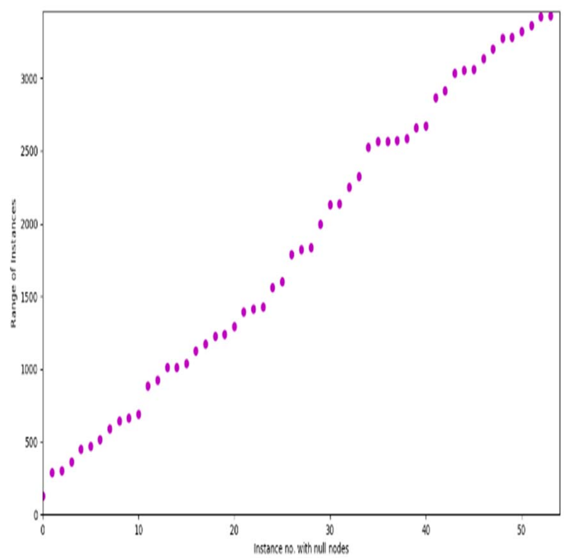

Figure 3: Scattered Graph plotted for determining instances with missing value.

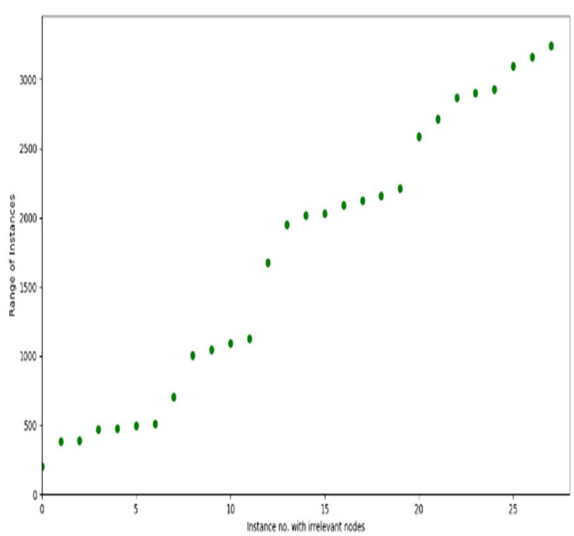

Figure 4: Scattered Graph plotted for determining instances with irrelevant values

B. Feature selection

In order to determine the features that are highly correlated the Pearson correlation technique is applied where the relation between the numerical attributes is found which is as shown in figure 5.

From the following plot it can be determined that the attribute 9 is highly correlated with attribute 7. Similarly attribute 12 and 10, 15 and 13, 18 and 16 are highly correlated. The correlation value ranges from -1 to 1. -1 represents high negative correlation, 0 represents no correlation while 1 represents high positive correlation. From the above plot it is clear that there exists a high positive correlation among the considered attributes.

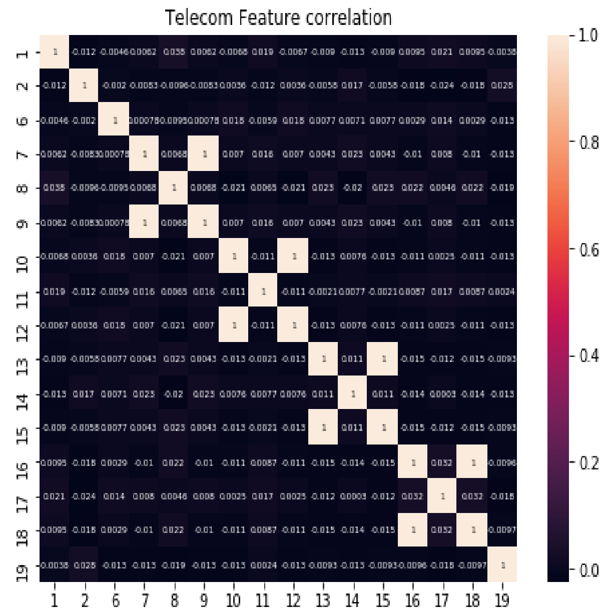

Figure 5: Graph plotted for correlation matrix

C. Feature Extraction

Feature Extraction is an attribute reduction process and this process results in a much smaller and richer set of attributes. It is found to provide the outcome which is as same as that of the data before the extraction. Conditions that have to be satisfied after feature extraction are:

- There should not be any linear relationships between the attributes.
- Every attribute should have a large variance.

The figure 6 depicts that on considering correlated attributes the dependency remains constant for each instance. Hence these attributes can be discarded resulting in feature extraction.

From the result it can be determined that for each instance the ratio between the charge and minute is same. Hence the charge attribute can be obtained for different time period as follows:

Total day charge = 0.17 * total day minutes (Eq.2)
Total eve charge = 0.085 * total eve minutes (Eq.3)
Total night charge = 0.045 * total night minutes (Eq.4)
intl charge = 0.27 * intl minutes (Eq.5)

```
Total day charge/minute: 0.1700113164843455
Total eve charge/minute: 0.08500506585612969
Total night charge/minute: 0.044993870044953005
Total international charge/minute: 0.27

Total day charge/minute: 0.16998762376237625
Total eve charge/minute: 0.08501278772378518
Total night charge/minute: 0.04500786163522012
Total international charge/minute: 0.27007299270072993

Total day charge/minute: 0.17000821692686935
Total eve charge/minute: 0.08498349834983498
Total night charge/minute: 0.04501845018450184 6
Total international charge/minute: 0.269672131147541

Total day charge/minute: 0.17000668002672012
Total eve charge/minute: 0.08497576736672051
Total night charge/minute: 0.04499746063991874
Total international charge/minute: 0.2696969696969697

Total day charge/minute: 0.17000599880023995
Total eve charge/minute: 0.08503034389750505
Total night charge/minute: 0.04499732477260567
Total international charge/minute: 0.2702970297029703

Total day charge/minute: 0.17000895255147716
Total eve charge/minute: 0.08499546690843156
Total night charge/minute: 0.04502206964198136
Total international charge/minute: 0.2698412698412698
```

Figure 6: The constant value obtained for different instances on considering the correlated attributes

As the charge attribute is a constant multiple of minute attribute, it can be eliminated from the dataset to reduce the dimensionality. Using the above four equations we can calculate the value of eliminated feature if required, by analysing the dependency between them.

D. Prediction

The RandomForest classifier is used for the prediction. It generates the subset of the given dataset and plots a decision tree for each subset. Based on the ranking the prediction is done with the label having the highest vote.The accuracy obtained for the dimensionality reduced dataset in our work is 92% which is greater than the accuracy obtained using other feature selection techniques that are used in our reference papers.

Accuracy: 0.9136690647482014

	Predicted False	Predicted True
Actual False	699	6
Actual True	66	63

Figure 7: Accuracy and confusion matrix of pre-processed data

Accuracy: 0.9218559218559218

	Predicted False	Predicted True
Actual False	705	4
Actual True	60	50

Figure 8: Accuracy and confusion matrix of extracted data

Figure 7 shows the accuracy and the confusion matrix of the pre-processed data which is comparatively less than that of the accuracy of extracted data given in figure 8.

E. ROC Curve

ROC (Receiver Operating Characteristics) curve is plotted by considering the True Positive rate (TPR) also called as recall or Sensitivity on Y-axis and False Positive rate (FPR) on X-axis.

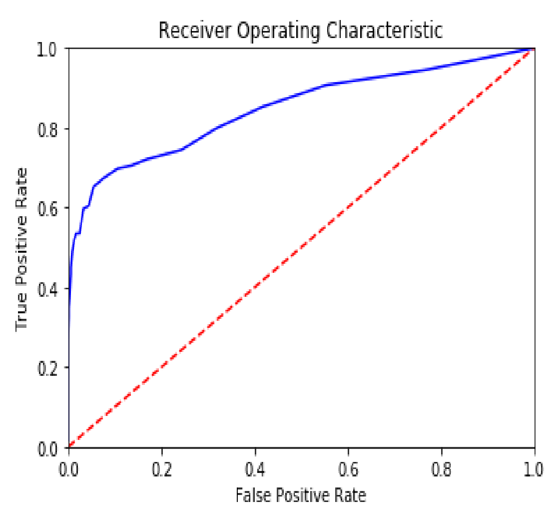

Figure 9: ROC curve plotted for pre-processed data

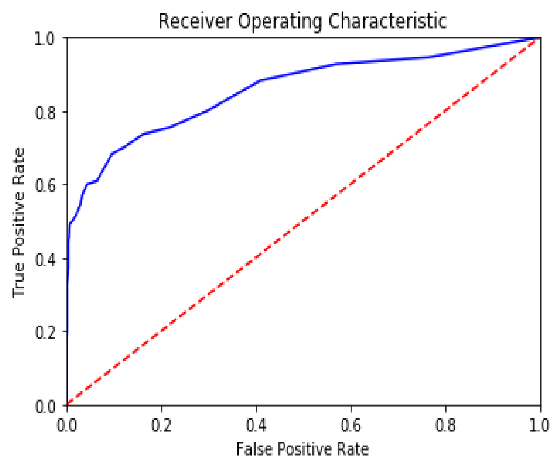

Figure 10: ROC curve plotted for extracted data

From the figure 9 and figure 10 it can be determined that the ROC curve plotted for pre-processed data is similar to that of the curve plotted for extracted data. From this it can be analyzed that the accuracy level of the extracted data is more by 1% to 2%, implying that the same prediction can be made with less size as the file size is reduced from 303 KB to 232 KB.

V. Conclusion

In telecom industry the prediction model must be efficient in addition to prediction accuracy. This efficiency is obtained by building the best training model which is reduced in terms of both size and dimensionality. The feature selection and extraction techniques used helps to obtain the efficient features that provides the accurate prediction. We can conclude that the prediction model built with reduced size of 232 KB is able to perform the tasks similar to or with more accuracy of 92% when compared to the original prediction model with size of 303 KB.

References

[1]. SyafiqahNafis, MokhairiMakhtar, Mohd Khalid Awang, M. NordinA.Rahman, Mustafa Mat Deris,"Feature Selection and classification model for customer churn", Journal of Theoretical and Applied Information Technology 31st May 2015, Volume 75, No. 3, ISSN:1992-8645

[2] Wei-Chao Lin,Chih-Fong Tsai, Shih-Wen Ke, "Dimensionlaity and data reduction in telecom churn prediction", Kybernetes Vol. 43, No. 5, 2014, pp.737-749.

[3]V. Umayaparvathi, K. Iyakutti, "Automated Feature Selection and Churn Prediction using Deep Learning Models", International Research Journal of Engineering and Technology (IRJET), Volume:04, Issue:03, Mar 2017.

[4] Varun E and Dr. PushpaRavikumar, "Attribute Selection for Telecommunication Churn Prediction", International Journal of Engineering & Technology,Vol. 7, No 4.39, 2018, pp.506-509.

[5] Varun E and Dr. PushpaRavikumar, "Community Mining In Multi-Relational and Heterogeneous Telecom Network", IEEE 6th International Conference on Advanced Computing (IACC-2016), DOI:10.1109/IACC.2016.15.

[6] Dr. M Balasubramanian, M Selvarani, "Churn Prediction in mobile telecom system using data mining techniques", published inInternational Journal of scientific and research publications. Issue 4, April 2014,Volume 4, ISSN 2250-3153.

[7] ManpreetKaur, Dr. PrernaMahajan, "Churn Prediction in Telecom Using R", International Journal of Engineering and Technical Research(IJER), ISSN: 2321-0869, Volume-3, Issue-5, May2015.

[8]. Pushpa, G Shoba, "Social Network Classifier for Churn Prediction in Telecom Data", 2013 International Conference on Advanced Computing and Communication Systems(ICACCS-2013), Dec.19-21, Coimbatore, INDIA.

[9]. N.Kamalraj Dr. A. Malathi, Department of computer science, Dr. SNS Rajalakshmi College of Arts and Science Government Arts College, "Applying Data Mining Techniques in Telecom Churn Prediction", International Journal of Advanced Research in CS and Software Engineering.

978-1-7281-3241-9/19/$31.00 © 2019 IEEE