International Journal of Engineering & Technology, 7 (4.39) (2018) 506-509

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Attribute Selection for Telecommunication Churn Prediction

Varun E [1],*, Dr. Pushpa Ravikumar [2]

[1]Assistant Professor, Department of Computer Science & Engineering, AIT College, India.
[2]Professor & Head, Department of Computer Science & Engineering, AIT College, India.
**Corresponding author E-mail: varune@aitckm.in*

Abstract

The telecommunication industries customer's bases are increasing every day. The industries are expected to significant loss of income due to increasing competition in drawing customers towards their customer bases. It is important to find the cause for losing customers and identifying the importance of the customer and retain them. The customer leaving the present telecom customer base and moving to other telecom service providers is called as churn. The telecommunication data set considered for identifying the importance of customer and churn prediction contain high dimensional data, it may contain redundant and inappropriate attributes. To apply the data mining tasks it is difficult to deal with high dimensional data and it leads to inappropriate predictions. To apply data mining task it is necessary to preprocess the data and reduce high dimensional data to low dimensional data without losing the prediction information. The reduced low dimensional data gives best results in churn prediction. This work focus on different attribute important measures and selection methods for identify the best subset of attributes for churn prediction. The experimental results of different attribute selection methods produces significant subset of attributes from high dimensional telecom dataset. The approach proven that it is helpful for predictive accuracy of further telecom churn management.

***Keywords**: Churn, Telecommunication Dataset, Attribute Selection, Prediction, Feature Selection.*

1. Introduction

The churning of customer's from telecom industries is one of the intensifying issues in rapidly growing and economical specific telecom industries. Telecom companies have shifted their view to acquire new customers into retain the current customers in their telecom base [1]. The retention of currently existing customers improved in financial growth and also reduced their marketing risks. Therefore the task of churn prediction is significantly intensifying as a part of telecom industries decision making and planning. The main objective of customer relationship management is to concentrate on customer retention. The significance of churn prediction has led to development of different tools and procedures that supports the task of classification and prediction [2].
The predictive classification and modeling requires attribute selection, since it is an important step to be carried out for data mining approach. Various attribute selection approach have been proposed example – Random Forest, stepwise selection, boruta etc. most of them have confirmed their importance for refining the predictive accuracy. All the attribute selection approaches classified into three main approaches are: Filter approach, Wrapper Approach and Embedded approach [3]. The same telecom attributes with the above three approach will produce different set of significant attributes. If the number of attributes is more than the optimal then the data mining approach may exhibit decrease in accuracy. Therefore it is important to select possibly small sub set of features form high dimensional data [4].
This paper concentrates mainly on analyzing the importance of attributes to build the churn prediction model accurately. The telecommunication dataset with 76 attributes are used for the feature extraction. The analysis is conducted on different attribute selection methods like random forest, boruta and stepwise forward selection to explore the most significant attributes.

2. Feature Selection Approach

Feature selection in data mining is the process of selecting important subset of attributes from the large dataset. The industry sectors these days generating large amount of high dimensional data day by day. Manual selection of the important attributes from large high dimensional data is difficult task [5]. The telecom data set used has contained multivariate attributes having significance and insignificant attributes [6]. To select the most important attributes, different attribute section approach is used. The three main approaches for feature selection in data mining are:

2.1. Filter Approach

The filter approach of feature selection uses single factor analysis technique. It examines the predictive power of each individual variable one by one. The data set containing larger number of attributes should use filter approach instead of brute force subset approaches. The predictive variable selection depends on the general characteristics of the training data set. The selected variables are independent of other models. Compared to wrapper approach, this approach is faster, computationally simple and scalable. This method uses mathematical evaluation function like Gain Ratio, correlation based techniques, chi-Square, information gain etc. that are based in internal characteristics of training data set. Since mathematical evaluation function is used we can know exactly why a given attribute is selected or not selected.

2.2. Wrapper Method

Wrapper method searches the optimal subset of features by using predictor or trained algorithms. This approach uses the different combinations or subsets of attributes from the dataset and finds the best subset of features. When using wrapper approach the user practically should consider (1) providing space of all possible combinations of subsets, and (2) How to evaluate the predictor performance. The best subset selection algorithm can be random forest, forward or backward selection etc. the wrapper approach considers the algorithm as a black box, which feeds all the attributes at once and the algorithm returns the subset of important attributes. The algorithm considers the interaction between the attributes it gives the result of subset by considering the relationship between the attributes. If the number of variables is not too large extensive search on all possible combinations can be done. One of the important packages available in R is Boruta package which is a wrapper around the random forest approach. the performance evaluation of the approach can be done by using cross validation method.

2.3. Embedded Method

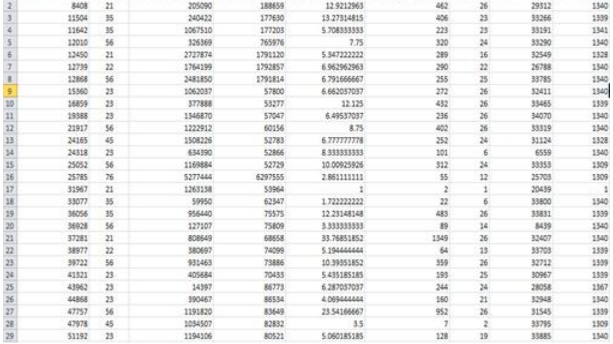
The embedded approach takes the advantages of both the wrapper and filter approach. This approach identifies the best features by using attribute subset and the performance of the model itself. In this approach the subset feature selection and the predictive model building cannot be separated. The commonly used embedded feature selection is regularization method. It performs analysis by statistical approach. Regularization approach introduces additional constraints for increasing performance of the prediction. Examples of regularization algorithms are the LASSO, Elastic Net

3. Methodology

3.1. Dataset

Data collection is the process of collecting information or high dimensional dataset on target attributes [7]. The dataset used for this work is telecommunication data with 76 different attributes of both prepaid and postpaid customers. The 30000 customers information of telecommunication interactions are considered for significant attribute selection approach. The service number calls are discarded during preprocessing. Some example attributes of telecom data set are cust_age, avg_call_cnt_per_day, ser_prepaid_avg_call_cnt_per_day, recharge_count etc.

The dataset tuples consisting of NA values can be identified by is.na function, the sum function returns total number of NAs in dataset. Dataset entry with NAs can be replaced by the mean value or calculating its value by deriving values from stored attributes. Example recharge_count can be derived from avg_call_cnt_per_day. The attribute mean is used for some of the attribute like customers cust_age, avg_call_cnt_per_day etc. The sample telecommunication dataset is shown in figure 1.



1	SUBSCRIPTIONKEY	cust_age	CUSTOMER_ACcnt_KEY	BILLING_ACcnt_KEY	avg_call_cnt_per_day	tot_call_cnt_to_month	day_count	GEOGRAPHY_KEY	LOCATION_KEY
2	8408	21	205090	188659	12.9212963	462	26	29312	1340
3	11504	35	240422	177630	13.27314815	406	23	33266	1339
4	11642	35	1067510	177203	5.708333333	223	23	33191	1341
5	12010	56	326369	765976	7.75	320	24	33290	1340
6	12450	21	2727874	1791120	5.347222222	289	16	32549	1328
7	12739	22	1764199	1792857	6.962962963	290	22	26788	1340
8	12868	56	2481850	1791814	6.791666667	255	25	33785	1340
9	15360	23	1062037	57800	6.662037037	272	26	32411	1340
10	16859	23	377888	53277	12.125	432	26	33465	1339
11	19388	23	1346870	57047	6.49537037	236	26	34070	1340
12	21917	56	1222912	60156	8.75	402	26	33319	1340
13	24165	45	1508226	52783	6.777777778	252	24	31124	1328
14	24318	23	634390	52866	8.333333333	101	6	6559	1340
15	25052	56	1169884	52729	10.00925926	312	24	33353	1309
16	25785	76	5277444	6297555	2.861111111	55	12	25703	1309
17	31967	21	1263138	53964	1	2	1	20439	1
18	33077	35	59950	62347	1.722222222	22	6	33800	1340
19	36056	35	956440	75575	12.23148148	483	26	33831	1339
20	36928	56	127107	75809	3.333333333	89	14	8439	1340
21	37281	21	808649	68658	33.76851852	1349	26	32407	1340
22	38977	22	380697	74099	5.194444444	64	13	33703	1339
23	39722	56	931463	73886	10.39351852	359	26	32712	1339
24	41321	23	405684	70433	5.435185185	193	25	30967	1339
25	43962	23	14397	86773	6.287037037	244	24	28058	1367
26	44868	23	390467	86534	4.069444444	160	21	32948	1340
27	47757	56	1191820	83649	23.54166667	952	26	31545	1339
28	47978	45	1034507	82832	3.5	7	2	33795	1309
29	51192	23	1194106	80521	5.060185185	128	19	33885	1340

Figure 1: Telecom Dataset

3.2. Feature Selection

Attribute selection is an important step for any model building approach. The next step after preprocessing the data is mining for important attributes. The significant attributes become the potential for customer churn prediction model. Particular for churn prediction, the are many categories of data including 1) customer care service details, 2) customer demography and personal details, 3) customer credit score, 4) bill and payment details, 5) customer usage pattern, and 6) customer value added services. There are several attributes that influence the customer to churn from one telecom base to another [8]. To identify such influencing attributes brute force method is used for exploratory analysis fo selecting top N-attributes which are dominant for customer churn. The brute force method for large number of attributes produce maximum subsets, which in-turn effects the processing overhead. The subset selection generates 2p different combinations of subset and it’s a processing overhead. The forward and backward selection generates less number of combinations as shown in formula 1.

$$\text{Number of subsets} = 1+P\,(P+1)/2 \qquad (1)$$

One of the best ways for implementing feature selection is Boruta package it uses wrapper method recursive feature elimination that finds the importance of a feature by creating shadow features. This paper focuses on different feature selection approach.

3.3. Framework

The strategy to find significant subset includes three dimensional approaches, which is evaluation criteria, processing feedback and well known algorithms [9]. Closed loop or wrapper approach is based on attribute selection which uses predictor performance. Wrapper approach gives solution for better attribute selection. The below algorithm depicts the principle of wrapper approach. In this framework a structure is defined for intelligent feature selection, which uses suitable feature selection algorithm in accordance with intended application.

Algorithm

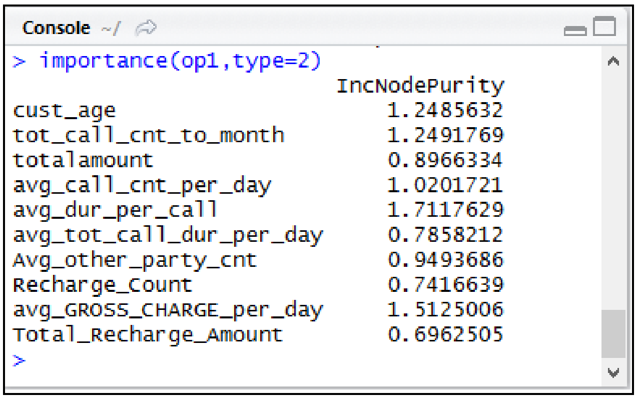
```
1.    Telecom dataset D with p labelled class or variables where v={a1 ,a2 ,….ap}
2.    Variable selection search – set j=1 select a distinct subset of variables Sj where 1<= Sj <=p.
1.    Induce learning algorithm.
2.    Evaluate the resulting model.
3.    Selected attributes
```

4. Results

4.1. Random Forest

Random Forests are similar to a famous Ensemble technique called Bagging. Random Forests correlates several trees which are generated by the different bootstrapped samples from training Data. And then we simply reduce the Variance in the Trees by averaging them. The idea is to build lots of Trees in such a way to make the Correlation between the Trees smaller.

Random forest algorithm is applied to the telecommunication dataset for feature selection. Consider to fit telecom dataset variables with 500 random forest trees, the algorithm returns the values of IncNodePurity as shown in figure 2. More useful variables achieve higher increases in node purities.



```
> importance(op1,type=2)
                         IncNodePurity
cust_age                     1.2485632
tot_call_cnt_to_month        1.2491769
totalamount                  0.8966334
avg_call_cnt_per_day         1.0201721
avg_dur_per_call             1.7117629
avg_tot_call_dur_per_day     0.7858212
Avg_other_party_cnt          0.9493686
Recharge_Count               0.7416639
avg_GROSS_CHARGE_per_day     1.5125006
Total_Recharge_Amount        0.6962505
>
```

Figure 2: Different IncNodePurity values

The variable importance plot as in figure 3 is obtained by growing some trees, then we can use simple functions importance(teledata) which represents the *mean increase in node purity*. It can be seen that recursive feature elimination algorithm has selected avg_dur_per_call , cust_age etc. as the important feature among the 76 features in the dataset.

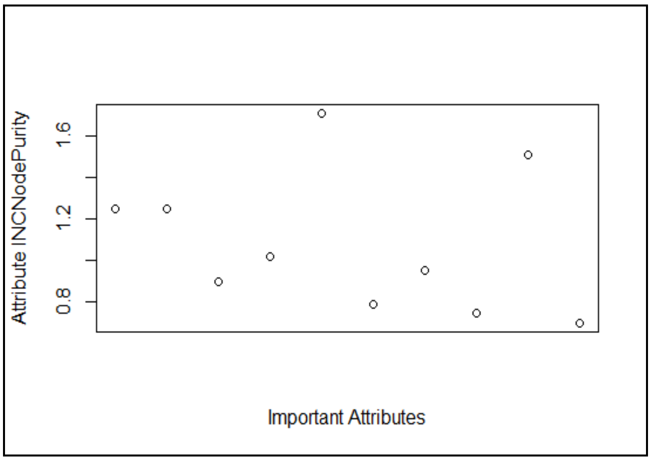


Figure 3: Attribute importance graph

The figure 4 plot shows the Error and the Number of Trees. It can be easily notice that the Error is dropping as adding more and more trees and averaging them.

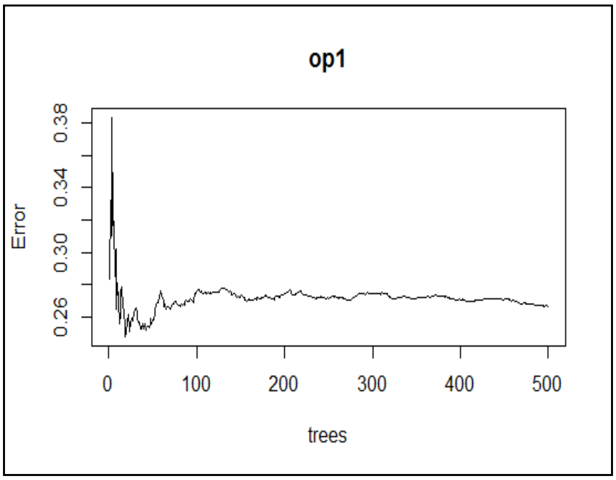
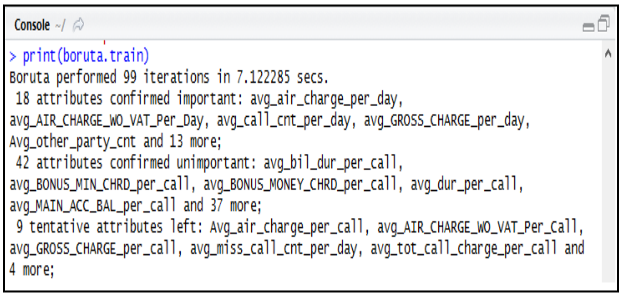


Figure 4: Error v/s Number of trees

4.2. Boruta Algorithm

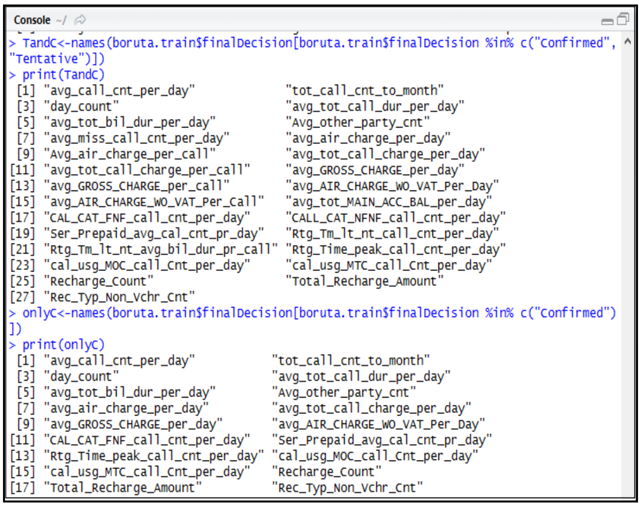
The figure 5 shows the variable importance result for the telecommunication dataset D. The algorithm performed 99 iterations for 76 different attributes, with that 18 attributes is conformed as important attributes, 42 other attributes are considered as unimportant and 9 attributes are considered as tentative attributes.



```
> print(boruta.train)
Boruta performed 99 iterations in 7.122285 secs.
 18 attributes confirmed important: avg_air_charge_per_day,
avg_AIR_CHARGE_WO_VAT_Per_Day, avg_call_cnt_per_day, avg_GROSS_CHARGE_per_day,
Avg_other_party_cnt and 13 more;
 42 attributes confirmed unimportant: avg_bil_dur_per_call,
avg_BONUS_MIN_CHRD_per_call, avg_BONUS_MONEY_CHRD_per_call, avg_dur_per_call,
avg_MAIN_ACC_BAL_per_call and 37 more;
 9 tentative attributes left: Avg_air_charge_per_call, avg_AIR_CHARGE_WO_VAT_Per_Call,
avg_GROSS_CHARGE_per_call, avg_miss_call_cnt_per_day, avg_tot_call_charge_per_call and
4 more;
```

Figure 5: output of variable selection using boruta

Boruta algorithm returns three different factors for attributes they are confirmed, rejected and tentative, which is the final result of feature selection. The figure 6 shows tentative and confirmed attributes for the telecom dataset D. Attributes that are considerably better then the tentative attributes are considered to be confirmed. The tentative attributes have importance consequently close to confirmed attributes but the algorithm not able to take decision with the default number of boruata algorithm run.



```
> TandC<-names(boruta.train$finalDecision[boruta.train$finalDecision %in% c("Confirmed",
"Tentative")])
> print(TandC)
 [1] "avg_call_cnt_per_day"             "tot_call_cnt_to_month"
 [3] "day_count"                        "avg_tot_call_dur_per_day"
 [5] "avg_tot_bil_dur_per_day"          "Avg_other_party_cnt"
 [7] "avg_miss_call_cnt_per_day"        "avg_air_charge_per_day"
 [9] "Avg_air_charge_per_call"          "avg_tot_call_charge_per_day"
[11] "avg_tot_call_charge_per_call"     "avg_GROSS_CHARGE_per_day"
[13] "avg_GROSS_CHARGE_per_call"        "avg_AIR_CHARGE_WO_VAT_Per_Day"
[15] "avg_AIR_CHARGE_WO_VAT_Per_Call"   "avg_tot_MAIN_ACC_BAL_per_day"
[17] "CAL_CAT_FNF_call_cnt_per_day"     "CALL_CAT_NFNF_call_cnt_per_day"
[19] "Ser_Prepaid_avg_cal_cnt_pr_day"   "Rtg_Tm_lt_nt_call_cnt_per_day"
[21] "Rtg_Tm_lt_nt_avg_bil_dur_pr_call" "Rtg_Time_peak_call_cnt_per_day"
[23] "cal_usg_MOC_call_Cnt_per_day"     "cal_usg_MTC_call_Cnt_per_day"
[25] "Recharge_Count"                   "Total_Recharge_Amount"
[27] "Rec_Typ_Non_Vchr_Cnt"
> onlyC<-names(boruta.train$finalDecision[boruta.train$finalDecision %in% c("Confirmed")
])
> print(onlyC)
 [1] "avg_call_cnt_per_day"           "tot_call_cnt_to_month"
 [3] "day_count"                      "avg_tot_call_dur_per_day"
 [5] "avg_tot_bil_dur_per_day"        "Avg_other_party_cnt"
 [7] "avg_air_charge_per_day"         "avg_tot_call_charge_per_day"
 [9] "avg_GROSS_CHARGE_per_day"       "avg_AIR_CHARGE_WO_VAT_Per_Day"
[11] "CAL_CAT_FNF_call_cnt_per_day"   "Ser_Prepaid_avg_cal_cnt_pr_day"
[13] "Rtg_Time_peak_call_cnt_per_day" "cal_usg_MOC_call_Cnt_per_day"
[15] "cal_usg_MTC_call_Cnt_per_day"   "Recharge_Count"
[17] "Total_Recharge_Amount"          "Rec_Typ_Non_Vchr_Cnt"
```

Figure 6: tentative and confirmed attributes

The graph in figure 7 shows the attribute importance structure of the telecommunication dataset. The graph represents the box plot which indicates the importance of each variable in the telecom dataset. The various colors in the graph represent the significance of the variables. Red boxplots represent minimal scores of the attributes. The yellow box plot represents the average score of the attribute. The green boxplots corresponds to the scores of confirmed attributes in the given data set respectively. According to variable importance graph, by using wrapper method 18 attributes came out as critical variables among the total of 76 attributes of telecommunication dataset. These 18 attributes are considered to have huge influence on telecom churn prediction model.

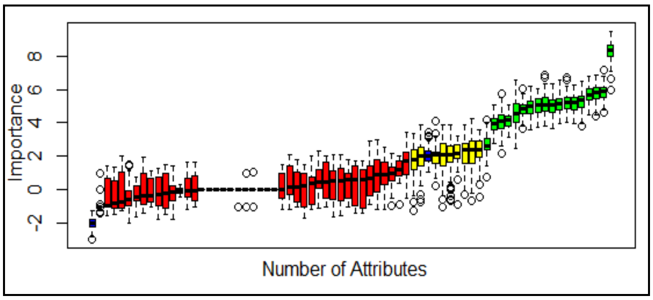
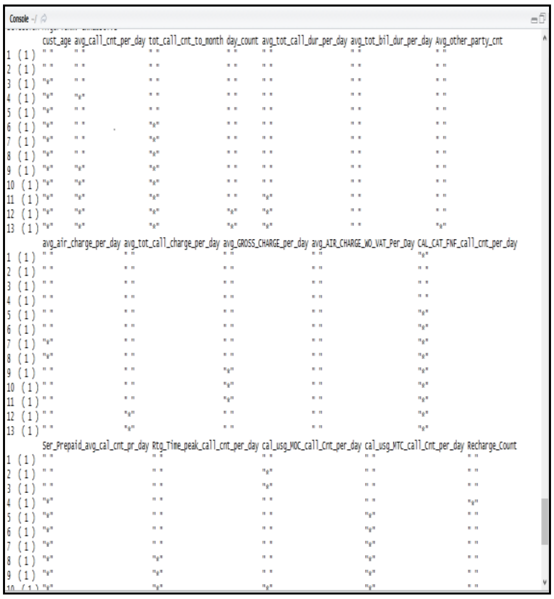


Figure 7: Box plot of attribute importance.

4.3. Stepwise Forward Selection

The stepwise regression approach for telecommunication dataset attribute selection is considered for selected attributes. The stepwise selection iteratively adds and removes the variable for variable selection model and finally returns the best performing model. There are two main approaches for stepwise regressing they are forward selection and back elimination.



```
        cust_age avg_call_cnt_per_day tot_call_cnt_to_month day_count avg_tot_call_dur_per_day avg_tot_bil_dur_per_day Avg_other_party_cnt
1  ( 1 ) " "      " "                  " "                   " "       " "                      " "                     " "
2  ( 1 ) " "      " "                  " "                   " "       " "                      " "                     " "
3  ( 1 ) "*"      " "                  " "                   " "       " "                      " "                     " "
4  ( 1 ) "*"      "*"                  " "                   " "       " "                      " "                     " "
5  ( 1 ) "*"      " "                  " "                   " "       " "                      " "                     " "
6  ( 1 ) "*"      " "                  "*"                   " "       " "                      " "                     " "
7  ( 1 ) "*"      " "                  "*"                   " "       " "                      " "                     " "
8  ( 1 ) "*"      " "                  "*"                   " "       " "                      " "                     " "
9  ( 1 ) "*"      "*"                  "*"                   " "       " "                      " "                     " "
10  ( 1 ) "*"     "*"                  "*"                   " "       " "                      " "                     " "
11  ( 1 ) "*"     "*"                  "*"                   " "       "*"                      " "                     " "
12  ( 1 ) "*"     "*"                  "*"                   "*"       "*"                      " "                     " "
13  ( 1 ) "*"     "*"                  "*"                   "*"       "*"                      " "                     "*"
        avg_air_charge_per_day avg_tot_call_charge_per_day avg_GROSS_CHARGE_per_day avg_AIR_CHARGE_WO_VAT_Per_Day CAL_CAT_FNF_call_cnt_per_day
1  ( 1 ) " "                    " "                         " "                      " "                           "*"
2  ( 1 ) " "                    " "                         " "                      " "                           " "
3  ( 1 ) " "                    " "                         " "                      " "                           " "
4  ( 1 ) " "                    " "                         " "                      " "                           " "
5  ( 1 ) " "                    " "                         " "                      " "                           "*"
6  ( 1 ) " "                    " "                         " "                      " "                           "*"
7  ( 1 ) "*"                    " "                         " "                      " "                           "*"
8  ( 1 ) "*"                    " "                         " "                      " "                           "*"
9  ( 1 ) " "                    " "                         "*"                      " "                           "*"
10  ( 1 ) " "                   " "                         "*"                      " "                           "*"
11  ( 1 ) " "                   " "                         "*"                      " "                           "*"
12  ( 1 ) " "                   "*"                         " "                      " "                           "*"
13  ( 1 ) " "                   "*"                         " "                      " "                           "*"
        Ser_Prepaid_avg_cal_cnt_pr_day Rtg_Time_peak_call_cnt_per_day cal_usg_MOC_call_Cnt_per_day cal_usg_MTC_call_Cnt_per_day Recharge_Count
1  ( 1 ) " "                            " "                            " "                          " "                          " "
2  ( 1 ) " "                            " "                            "*"                          " "                          " "
3  ( 1 ) " "                            " "                            "*"                          " "                          " "
4  ( 1 ) "*"                            " "                            " "                          " "                          "*"
5  ( 1 ) "*"                            " "                            " "                          "*"                          " "
6  ( 1 ) "*"                            " "                            " "                          "*"                          " "
7  ( 1 ) "*"                            " "                            " "                          "*"                          " "
8  ( 1 ) "*"                            "*"                            " "                          "*"                          " "
9  ( 1 ) "*"                            "*"                            " "                          "*"                          " "
10  ( 1 ) "*"                           "*"                            "*"                          "*"                          " "
```

Figure 8: stepwise regression- forward selection

The forward selection starts with single attribute from the dataset, iteratively adds the most significant attributes and stops if no improvement in the selected subset. The backward selection starts with all variables and iteratively removes the least significant attributes. The asterisk in figure 8 specifies that the particular variable is included in the corresponding model. It can be realized that the best 4-variable model consists of cust_age, avg_call_cnt_per_day, ser_prepaid_avg_call_cnt_per_day and recharge_count.

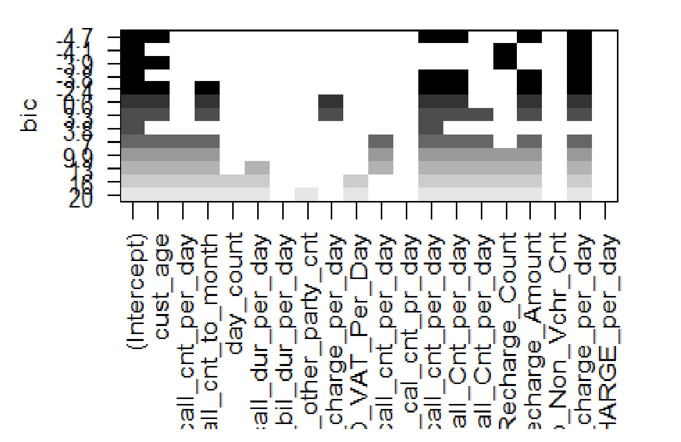


Figure 9: BIC for telecom data attribute selection

The figure 9 illustrations the BIC plot which shows the best subsets for telecom data mining, the subset include cust_age, Recharge_amount, avg_call_cnt_per_day and avg_Gross_charge_per_day. The vertical axis shows the drop in BIC of unimportant attributes compared to important attributes. The plot in the graph supports multiple candidate models with the BIC on Y axis. Lower value of BIC is better which is depicted at the top of the graph. The graph has -4.7 as the lower value of BIC and 20 as the higher value of BiC.

5. Conclusion

The proposed work has focused on data mining techniques and R packages to perform comparative study on feature extraction of telecom dataset. The numbers of different features are classified based on the selection framework. The R tool provides unique feature visualization with the help of graphs. The results include cust_age, Recharge_amount, avg_call_cnt_per_day and avg_Gross_charge_per_day as the most important attributes. The results acquired from the attribute selection approaches can be further used with domain intelligence to obtain additional specific attributes which helps for worthy prediction. The final attributes selection can be used to build the social telecom network. The proposed framework is useful for significant attribute selection and to build the accurate telecom churn prediction model.

References

[1] Ammar A Ahmed, Dr. D. Maheswari linen, “A Review And Analysis Of Churn Prediction Methods For Customer Retention In Telecom Industries”, in Proc. IEEE International Conference on Advanced Computing and Communication Systems (ICACCS -2017), January, pp: 06 – 07, 2017, Coimbatore, India.

[2] Sepideh Hassankhani Dolatabadi, Farshid Keynia, “Designing of customer and employee churn prediction model based on data mining method and neural predictor”, in Proc. IEEE 2nd International Conference on Computer and Communication Systems (ICCCS), pp: 74 – 77, 2017.

[3] Deepshika Nagpal, Rashi Srivastava, Deepti Mehrotra, Anuranjana “Feature Selection Approach for Reducing the Power Consumption for a Greener Environment” 2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017).

[4] D.P. Acharjya, T.K. Das “A framework for attribute selection in marketing using rough computing and formal concept analysis” 2017 Production and hosting by Elsevier Ltd on behalf of Indian Institute of Management Bangalore.

[5] [5] Cong Jin, Shu-Wei Jin, Li-Na Qin, “Attribute selection method based on a hybrid BPNN and PSO algorithms” Applied Soft Computing 12 (2012) 2147–2155 hosting by Elsevier Ltd.

[6] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, V. A. Kanade, “Analysis of customer churn prediction in telecom industry using decision trees and logistic regression”, in Proc. IEEE Symposium on Colossal Data Analysis and Networking (CDAN), pp: 1 – 4, 2016.

[7] Hui Li, Deliang Yang, Lingling Yang, YaoLu, Xiaola Lin, “Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction”, in Proc. IEEE International Conferences on Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications, pp: 163 – 169, 2016.

[8] Nishant Borude, Chandrakant Maher, Vishal Sarda, Aparna Santra, “Generic binary classifier tool for diagnosis of patients suffering from brain disorders in R”, in Proc. IEEE International Conference on Computing, Analytics and Security Trends (CAST), pp: 173 – 178, 2016.

[9] Sebastian Robitzsch, Faisal Zaman, Zhiguo Qu, John Keeney; Sven van der Meer, Gabriel-Miro Muntean, “E-stream: Towards pattern centric network incident discovery and corrective action recommendation in telecommunication networks”, in Proc. IFIP/IEEE International Symposium on Integrated Network Management (IM), pp: 842 – 845, 2015.