International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJEMR Transactions, online available on 1st Jun 2019. Link :http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-06

Title: DIGITAL PRODUCT SUBSCRIBERS CHURN PREDICTION MODEL USING LOGISTIC REGRESSION

Volume 08, Issue 06, Pages: 1–7.

Paper Authors

ANUP.B.A, JANAKI.K, BHOOMIKA.L, DEVI.B.S, MANISHA.S

RajaRajeswari College Of Engineering Bengaluru-560074

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

DIGITAL PRODUCT SUBSCRIBERS CHURN PREDICTION MODEL USING LOGISTIC REGRESSION

[1]ANUP.B.A, [2]JANAKI.K, [3]BHOOMIKA.L, [4]DEVI.B.S, [5]MANISHA.S
Department of Computer Science and Engineering
RajaRajeswari College Of Engineering Bengaluru-560074
anupba72@gmail.com, karur.janaki@gmail.com, bhoomikanth@gmail.com

ABSTRACTCustomer Churn Prediction is an important factor for any Organization. Customer churn directly affects the revenue and also depletes the growth of an Organization. Therefore, necessary action must be taken to avoid churning. Our aim is to build an efficient model using logistic regression which predicts the churners and non-churners for subscription based Digital Products. The accuracy of the model predicted is 81% which is satisfactory and effective. The churners are sent an email with exciting offers and discounts, which make them to continue with the Digital Product..

KEYWORDS
Churn Prediction, Data Pre-processing, Logistic Regression, Email Service.

1. INTRODUCTION

Nowadays, people rely mostly on Digital Products. The increased usage of digital products leads to the growth of company. The companies strive hard for fulfilling the demands of the customer and to survive in this competitive world. Losing a customer will be a loss to the company and also it leads to negative impact of the company The main focus of the company will be on how to predict and prevent the customers from churning.. As a result, we develop a model which predicts churners and non-churners and thus helps the company. Churners are the one who do not continue using the digital products and non-churners are the one who continue. To reduce the rate of churning, an email is sent to the customers that contains special offers and discounts.

Churn Prediction: Churn prediction is a process of analysing and predicting the subscribers or customers who end up using the Digital Product.

In this paper, the digital products like Literacy Pro (LP) and Literacy Pro Library (LPL) are considered. The licenses are provided to the schools where students use the licenses to access these digital products. The licenses are subscribed annually. Literacy Pro (LP) is a product where students take up test to check their English proficiency and Literacy Pro Library (LPL) is used to read books and take up quiz. In order to develop a predicting model, the supervised learning algorithm that is logistic regression algorithm is used to predict churners and non-churners. Simple Mail Transfer Protocol(SMTP) is a communication protocol used for email transmission. The churners are sent an email using this protocol.

In section 2, we have summarized some prior study of churn prediction of different algorithms. In section 3, we explain the methodology and the steps involved. Section 4 briefs about the result obtained. Section 5 describes the conclusion and future work of this project.

2. RELATED WORK

Decision tree is the straightforwardness technique for interpretation of the discovered rules. This algorithm will construct a tree with the training set and this includes a node. This node is an attribute and the branches are the corresponding attribute values. The disadvantage of this algorithm is there will be small change in the training data and this will lead to large variations and it would be an unstable algorithm[1]. J. Burez et al, by balancing the data it helps to increase churn prediction accuracy, it also insists significance of having managed data to predict churning of the customer. Weight random forest for a cost sensitive learner is used. But these random sampling method can decrease the efficiency of the data[2].

VeronikhaEffendy et al proposed a technique for handling the problem of imbalance data to improve the customer churn prediction. The proposed technique is a combination of sampling and Weighted Random Forest(WRF) to balance the data for the accuracy of churn prediction. The process involves sampling of imbalanced data problem and WRF will classify the data for accurate churn prediction. Combined sampling process will increase the accuracy values for accurate prediction of the data[3]. Ning Lu et al proposed a churn prediction model for telecom industry. Logistic regression and Gentle AdaBoost both these tecniques have certain limitations, no consideration of class scarcity and inability to finalize the reason for churn prediction[4].

Xiaojun Wu et al proposed an expectation strategy on SMOTE (Synthetic Minority Oversampling Technique) and AdaBoost for anticipating stir customer for web based business. The improved SMOTE is connected to process the unequal datasets by the blend of over-examining and under-inspecting. From the learning algorithm AdaBoost is utilized for adjusted datasets for frail classifier to arrange the expectation of the clients agitate [5].G. Ganesh Sundarkumar et al proposed oneclass SVM for improving the churn prediction and insurance fraud detection based on under-sampling. The data is undersampled using one-class SVM and then the classification is performed using some machine learning algorithms. Based on the result the Decision tree performs better than other classification algorithms and one-class SVM will reduce the system complexity and it will improve the prediction accuracy[6].

QiuhuaShen et al, proposed a framework for the betterment of churn prediction rate by the complementary fusion of multilayer features. The proposed framework will include the feature factorisation and feature is constructed for combination of features. Solving the unbalanced data problem, it enhances the churn prediction accuracy[7]. Aimee Backiel et al, proposed two element models to identify distinctive arrangement of churners by utilizing the blend of nearby and social highlights for stir expectation. An aggregate methodology for joining of two highlights. The customer information and social information from a cell phone specialist is used for testing. The proposed model contains the spreading actuation calculation which will spread the neighborhood and social factors among the social and nearby model and an aggregate model for brushing the highlights together. The result of assessment infers that the beat forecast is improved by utilizing a consolidated model of highlights as opposed to utilizing singular models or the spread element models. The disadvantages of this methodology is that the oversight of non-customer hubs in the production of call chart because of bigger volume of information diminishes the viability of agitate expectation. Another methodology is the exclusion of negative energies from the informal community. Informal community examination can upgrade the customer beat expectation[8].

As an undeniably related research, Tangetal coordinated an examination where they used measurement features, large scale monetary variables, and financial information, for instance, approach purchase inorder to envision customer shake decision. Diverged from these space related highlights, we consider the dynamic spatio-temporal"patterns"(diversity, dedication, normality) of spending exercises, and entropy offund trade and purchase trades. Interms of forecast, Tangetal changed over the first financial features into the decided features by applying the symmetrical polynomial gauge approach. In examination, we made novel features subject to spatio-common and choice practices of clients ,and developed a RandomForest model which is a social affair procedure for non-direct treeclassifiers. The model is fit for considering distinctive mixes of different features, and the most extreme profundity hyper-parameter relates to the level of polynomial terms. From this vantage point, our forecast strategy is like Tangetal as far as fusing higher-degree highlight combinations[9].

AmjadHudaib et al, proposed three half and half information digging models for the stir forecast application. The half and half models are created dependent on two stages to be specific the grouping stage for client information separating and the expectation stage for foreseeing client conduct. The principal model uses k-implies calculation for information sifting while Multilayer Perceptron Artificial Neural Networks (MLP-ANN) is utilized for agitate forecast. The second cross breed model utilizes various leveled bunching alongside MLP-ANN for expectation. The third model utilizes self-arranging maps (SOM) with MLP-ANN. On assessment it is discovered that the half and half models beat the single models while the principal cross breed model gives higher forecast exactness than the other two models. The confinement with this methodology is that the MLP-ANN devours practically equivalent handling time for extensive just as littler datasets[10].

3. PROPOSED METHOD

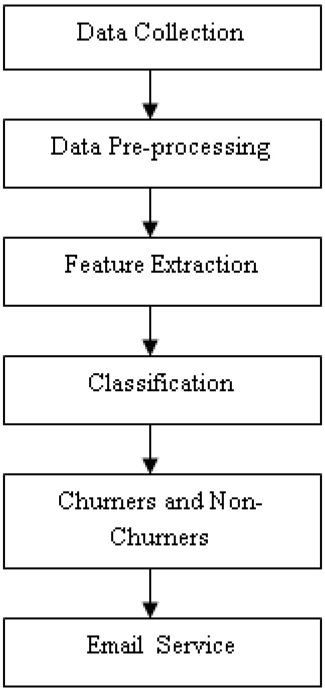


Fig 1: System Architecture

Fig 1 represents the system architecture of the project Digital Product Subscribers Churn Prediction Model Using Machine Learning.

The first step is the collection of data from various sources. The collected data is a raw data so that it should be pre-processed to get refined data. Data pre-processing removes the noise i.e., the unwanted data, null and NaN data. The third step is the feature extraction, where the feature is extracted from the pre-processed data. After extraction, the features are fed into Classification algorithm (Logistic Regression) where the data is divided into train and test data. This algorithm classifies Churners and Non-churners. Finally, the Churners are notified with an email with offers and discounts.

A. Data Collection

We collect three years of historical data (2016, 2017 & 2018) of subscribers, which represents the number of Literacy Pro (LP) licenses and Literacy Pro Library (LPL) licenses issued to them, number of quiz and test attempted per license. The School Data Set collected involves the fields like Customer, OrgID, LP Licenses of 2016, LPL Licenses of 2016, LP+LPL Licenses of 2016, Only LP Licenses, LP Licenses of 2017, LPL Licenses of 2017, LP+LPL Licenses of 2017, LP Licenses of 2018, LPL Licenses of 2018, LP+LPL Licenses of 2018.The test dataset contains fields like 2016_ID, Renewed_2017, 2016test_created_year,2016test_created_month, 2016_Test_Successful and similarly for the year 2017 and 2018.The Quiz dataset contains fields like org_id-2016, renewed-in-2017-flag, school_year-2016, quiz-month-2016, quiz-attempt-count-2016 and similarly for the year 2017 and 2018.

B. Data Pre-processing

Data Pre-processing is a technique that converts raw data set into a clean data set. It is a process where noise, null values, inconsistent values and unwanted values are removed. Pre-processing is an important step to be taken before the execution of the project. We combined all the features of school dataset, Test dataset and quiz dataset. The dataset consisted of null and NaN, which was not useful in developing the model and hence it was removed. The dataset also consisted of missing values. Example: The Organization Ids of the schools were present in the School dataset, but the test and quiz data were missing. This was refilled with the zero value.

C. Feature Extraction

After Data Pre-Processing, we obtain the cleansed data set for training the model. From these dataset, we derive parameters like rate of increase of license in 2017, rate of increase of license in 2018, ratio of test successful per license and ratio of quiz attempted per license. Based on these parameters we classify the subscribers as churners and non-churners.

D. Classification

Classification is a Supervised learning technique. Basically supervised learning technique is where the machine is trained first with some dataset which has correct output. After that, the machine is given with the new dataset that has to be tested. The supervised learning algorithm analyses the training dataset and produces the output for the new dataset. We use Logistic Regression for classification.

Logistic Regression: Logistic regression is a supervised classification algorithm, where the machine is trained first with training dataset and then the testing is done using test dataset. In the problem of classification, the output variable or the target variable which is also categorical: y, takes only the discrete values for set of inputs: X. In this paper, Binomial Logical Regression is considered. It is the target variable that has only 2 possible outputs (0 or 1) which may also be represented as “churn or not churn” respectively.

Logistic Regression is given by:
$\ln[p/(1-p)] = \alpha + \beta X + e$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"
- $\beta$is a regression Coefficient
- $\alpha$is an auxiliary value
- e is a constant

Logistic distribution constraints, the analyzed probability should always lie between 0 and 1.The estimated probability is calculated by using the formula:

$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

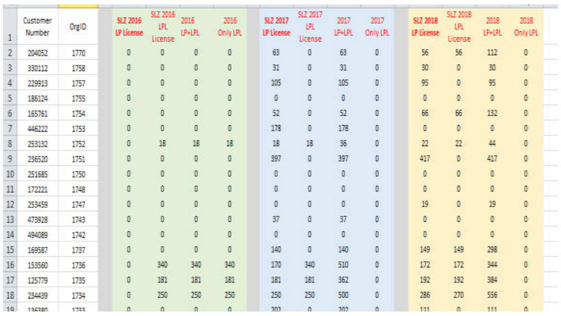
The model is trained with the extracted features like, rate of increase of license in the year 2017, rate of increase of license in the year 2018, ratio of quiz attempted and ratio of test successful of 2016, 2017 and 2018. After training, the model is tested with the new datasets which predicts the output (churners and non-churners).

E. Email Service

After predicting the output Churners and non-churners, the churners are sent with an email that uses Simple Mail Transfer Protocol (SMTP) protocol. The email consists of offers and discounts, so that the churners can continue using the digital products, which will be helpful for the organization. SMTP is a communication protocol which is used for delivering the mails. Simple Mail Transfer Protocol (SMTP) supports text, audio, image format.

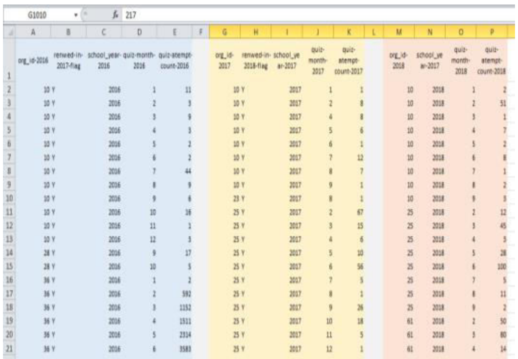
4. RESULT

Fig 2, Fig 3 and Fig 4 represents the data collected. Basically the collected data are unstructured and unformatted. Fig 2 represents the Number of Literacy Pro (LP) and Literacy Pro Library (LPL) licenses of the organization. Fig 3 represents the number of Quiz attempted per license. Fig 4 represents the number of Test attempted per license.



	Customer Number	OrgID	SLZ 2016 LP License	SLZ 2016 LPL License	2016 LP+LPL	2016 Only LPL	SLZ 2017 LP License	SLZ 2017 LPL License	2017 LP+LPL	2017 Only LPL	SLZ 2018 LP License	SLZ 2018 LPL License	2018 LP+LPL	2018 Only LPL
1														
2	204052	1770	0	0	0	0	63	0	63	0	56	56	112	0
3	330112	1758	0	0	0	0	31	0	31	0	30	0	30	0
4	229913	1757	0	0	0	0	105	0	105	0	95	0	95	0
5	186124	1755	0	0	0	0	0	0	0	0	0	0	0	0
6	165761	1754	0	0	0	0	52	0	52	0	66	66	132	0
7	446222	1753	0	0	0	0	178	0	178	0	0	0	0	0
8	253132	1752	0	18	18	18	18	18	36	0	22	22	44	0
9	236520	1751	0	0	0	0	397	0	397	0	417	0	417	0
10	251685	1750	0	0	0	0	0	0	0	0	0	0	0	0
11	172221	1748	0	0	0	0	0	0	0	0	0	0	0	0
12	253459	1747	0	0	0	0	0	0	0	0	19	0	19	0
13	473928	1743	0	0	0	0	37	0	37	0	0	0	0	0
14	494089	1742	0	0	0	0	0	0	0	0	0	0	0	0
15	169587	1737	0	0	0	0	140	0	140	0	149	149	298	0
16	153560	1736	0	340	340	340	170	340	510	0	172	172	344	0
17	125779	1735	0	181	181	181	181	181	362	0	192	192	384	0
18	234439	1734	0	250	250	250	250	250	500	0	286	270	556	0
19	136380	1733	0	0	0	0	202	0	202	0	111	0	111	0

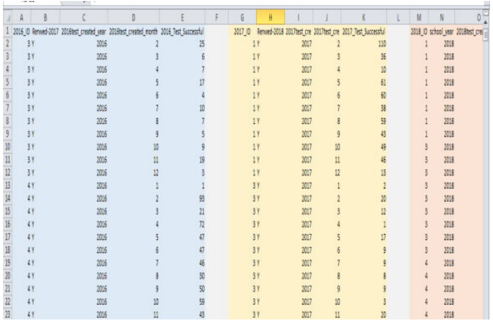
Fig 2: Number of Literacy Pro (LP) and Literacy Pro Library (LPL) Licenses of the organizations.



G1010 fx 217

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	org_id-2016	renwed-in-2017-flag	school_year-2016	quiz-month-2016	quiz-atempt-count-2016		org_id-2017	renwed-in-2018-flag	school_year-2017	quiz-month-2017	quiz-atempt-count-2017		org_id-2018	school_year-2017	quiz-month-2018	quiz-atempt-count-2018
2	10	Y	2016	1	11		10	Y	2017	1	1		10	2018	1	2
3	10	Y	2016	2	3		10	Y	2017	2	8		10	2018	2	51
4	10	Y	2016	3	9		10	Y	2017	4	8		10	2018	3	1
5	10	Y	2016	4	3		10	Y	2017	5	6		10	2018	4	7
6	10	Y	2016	5	2		10	Y	2017	6	1		10	2018	5	2
7	10	Y	2016	6	2		10	Y	2017	7	12		10	2018	6	8
8	10	Y	2016	7	44		10	Y	2017	8	7		10	2018	7	1
9	10	Y	2016	8	9		10	Y	2017	9	1		10	2018	8	2
10	10	Y	2016	9	6		23	Y	2017	8	1		10	2018	9	3
11	10	Y	2016	10	16		25	Y	2017	2	67		25	2018	2	12
12	10	Y	2016	11	1		25	Y	2017	3	15		25	2018	3	45
13	10	Y	2016	12	3		25	Y	2017	4	6		25	2018	4	3
14	28	Y	2016	9	17		25	Y	2017	5	10		25	2018	5	28
15	28	Y	2016	10	5		25	Y	2017	6	56		25	2018	6	100
16	36	Y	2016	1	2		25	Y	2017	7	5		25	2018	7	5
17	36	Y	2016	2	592		25	Y	2017	8	1		25	2018	8	11
18	36	Y	2016	3	1152		25	Y	2017	9	26		25	2018	9	2
19	36	Y	2016	4	1511		25	Y	2017	10	18		61	2018	2	50
20	36	Y	2016	5	2314		25	Y	2017	11	5		61	2018	3	80
21	36	Y	2016	6	3583		25	Y	2017	12	1		61	2018	4	14

Fig 3: Number of Quiz attempted per license



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	2016_ID	Renwed-2017	2016test_created_year	2016test_created_month	2016_Test_Successful		2017_ID	Renwed-2018	2017test_cre	2017test_cre	2017_Test_Successful		2018_ID	school_year	2018test_cre
2	3	Y	2016	2	25		1	Y	2017	2	110		1	2018	
3	3	Y	2016	3	6		1	Y	2017	3	36		1	2018	
4	3	Y	2016	4	7		1	Y	2017	4	10		1	2018	
5	3	Y	2016	5	17		1	Y	2017	5	61		1	2018	
6	3	Y	2016	6	4		1	Y	2017	6	60		1	2018	
7	3	Y	2016	7	10		1	Y	2017	7	38		1	2018	
8	3	Y	2016	8	7		1	Y	2017	8	59		1	2018	
9	3	Y	2016	9	5		1	Y	2017	9	43		1	2018	
10	3	Y	2016	10	9		1	Y	2017	10	49		3	2018	
11	3	Y	2016	11	19		1	Y	2017	11	46		3	2018	
12	3	Y	2016	12	3		1	Y	2017	12	13		3	2018	
13	4	Y	2016	1	1		3	Y	2017	1	2		3	2018	
14	4	Y	2016	2	93		3	Y	2017	2	20		3	2018	
15	4	Y	2016	3	21		3	Y	2017	3	12		3	2018	
16	4	Y	2016	4	72		3	Y	2017	4	1		3	2018	
17	4	Y	2016	5	47		3	Y	2017	5	17		3	2018	
18	4	Y	2016	6	47		3	Y	2017	6	9		3	2018	
19	4	Y	2016	7	46		3	Y	2017	7	9		4	2018	
20	4	Y	2016	8	30		3	Y	2017	8	8		4	2018	
21	4	Y	2016	9	50		3	Y	2017	9	9		4	2018	
22	4	Y	2016	10	59		3	Y	2017	10	3		4	2018	
23	4	Y	2016	11	43		3	Y	2017	11	20		4	2018	

Fig 4: Number of Test attempted per license

Fig 5 represents the Total number of Licenses purchased by the Organizations. x-axis indicates Organization ID and y-axis indicates Number of Licenses.

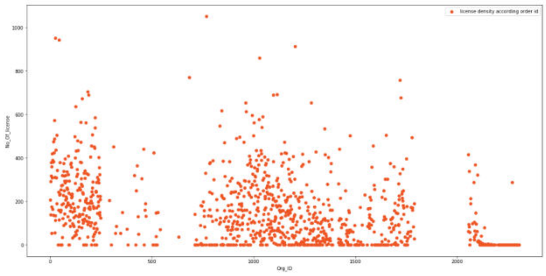
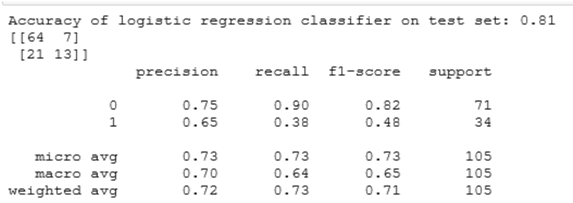


Fig 5: Total number of Licenses of Organizations

The performance of the classification model for the test data is indicated by Confusion Matrix. It also shows the performance of the classification algorithm. The accuracy of the model is around 81% which is satisfactory and effective. The precision, recall, F1-score and support is shown in the Fig 6. Fig 7 represents the Receiver Operating Characteristic (ROC) curve of the model. In Fig 7, x-axis represents the False Positive Rate and y-axis represents the True Positive Rate. The area under curve for 70% train data and 30% test data is observed as 0.73.



```
Accuracy of logistic regression classifier on test set: 0.81
[[64  7]
 [21 13]]
             precision    recall  f1-score   support

          0       0.75      0.90      0.82        71
          1       0.65      0.38      0.48        34

  micro avg       0.73      0.73      0.73       105
  macro avg       0.70      0.64      0.65       105
weighted avg       0.72      0.73      0.71       105
```

Fig 6 Confusion Matrix

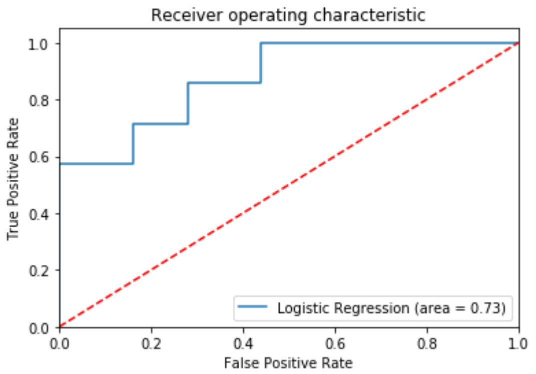


Fig 7: Receiver Operating Characteristic curve

Fig 8 represents the graph of Churners and Non-Churners. x-axis represents Organisation ID and y-axis indicates possible outcome. The Churners are indicated by ‘0’ and the Non-Churners are represented by ‘1’.

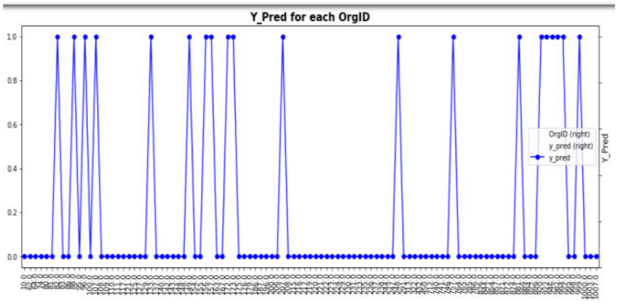


Fig 8: Representation of Churners and Non-churners

5. CONCLUSION

In this paper, we predicted which customers will continue using the product and who will not in future. Customer churn is one of the major problems that the companies are facing now. To solve this problem, the datasets are subjected for logistic regression algorithm. There are many machine learning algorithms that are used to predict the churners and n0n-churners. By using machine learning algorithms most accurate results of churn prediction is obtained. Initially, there are huge amount of data, these data are pre-processed to remove the outliers. The dataset consists the usage of the product for three years, based on the data of these three years, the prediction is made for the next upcoming years. These results will assist by identifying the churners who are at risk of ending up using the Digital Products and the necessary actions can be taken to retain the churners for continuing using the product. The result of this paper gives 81% of accuracy which is satisfactory and effective. It can be concluded that the most accurate result of churn prediction is obtained.

6. REFERENCE

[1]. Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification.John Wiley & Sons, 2012.

[2]. J. Burez and Dirk Van den Poel. "Handling class imbalance in customer churn prediction." Expert Systems with Applications, vol. 36, no. 3, pp. 4626-4636, 2009

[3]. VeronikhaEffendy and ZK AbdurahmanBaizal. "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest." In 2014 2nd International Conference on Information and Communication Technology (ICoICT), pp. 325-330.IEEE, 2014.

[4]. Ning Lu, Hua Lin, Jie Lu, and Guangquan Zhang. "A customer churn prediction model in telecom industry using boosting." IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pp. 1659-1665, 2014.

[5]. Xiaojun Wu and SufangMeng. "E-commerce customer churn prediction based on improved SMOTE and AdaBoost." In 2016 13th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-5.IEEE, 2016.

[6]. G. Ganesh Sundarkumar, Vadlamani Ravi, and V. Siddeshwar. "One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection." In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-7.IEEE, 2015.

[7]. QiuhuaShen, Hong Li, Qin Liao, Wei Zhang, and KoneKalilou. "Improving churn prediction in telecommunications using complementary fusion of multilayer features based on factorization and construction." The 26th Chinese Control and Decision Conference (2014 CCDC), pp. 2250-2255. IEEE, 2014.

[8]. AiméeBackiel, YannickVerbinnen, Bart Baesens, and GerdaClaeskens. "Combining local and social network classifiers to improve churn prediction." In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 651-658. IEEE, 2015.

[9]. TangL, ThomasL, FletcherM, PanJ, MarshallA(2014) Assessing the impact of derived behaviour informationon customer attrition in the financial service industry. EurJOperRes236(2):624–633

[10]. AmjadHudaib, RehamDannoun, Osama Harfoushi, RubaObiedat, and HossamFaris. "Hybrid Data Mining Models for Predicting Customer Churn."International Journal of Communications, Network and System Sciences, vol. 8, no. 05, pp. 91, 2015.