I.J. Modern Education and Computer Science, **2019, 6, 1-7**
Published Online June 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2019.06.01

Combined Appetency and Upselling Prediction Scheme in Telecommunication Sector Using Support Vector Machines

Lian-Ying Zhou
School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, China
Email: zhouly@ujs.edu.cn

Daniel M. Amoh, Louis K. Boateng and Andrews A. Okine
Jiangsu University/School of Computer Science and Communication Engineering, Zhenjiang, 212013, China
Email: {mdamoh@yahoo.com. quamelouis18@gmail.com. andyokine101@gmail.com}

Received: 04 April 2019; Accepted: 23 April 2019; Published: 08 June 2019

Abstract—Customer Relations Management (CRM) is an essential marketing approach which telecommunication companies use to interact with current and prospective customers. In recent years, researchers and practitioners have investigated customer churn prediction (CCP) as a CRM approach to differentiate churn from non-churn customers. CCP helps businesses to design better retention measures to retain and attract customers. However, a review of the telecommunication sector revealed little to no research works on appetency (i.e. customers likely to purchase new product) and up-selling (i.e. customers likely to buy upgrades) customers. In this paper, a novel up-selling and appetency prediction scheme is presented based on support vector machine (SVM) algorithm using linear and polynomial kernel functions. This study also investigated how using different sample sizes (i.e. training to test sets) impacted the classification performance. Our findings demonstrated that the polynomial kernel function obtained the highest accuracy and the least minimum error in the first three sample sizes (i.e. 80:20, 77:23, 75:25) %. The proposed model is effective in predicting appetency and up-sell customers from a publicly available dataset.

Index Terms—Customer Relations Management, Telecommunication, Churn prediction, Appetency prediction, Up-selling prediction, Support Vector Machines, classification.

I. Introduction

The rapidly mounting issues in the telecommunication sector today are growth and competition. Understanding customers is vital in the extremely competitive telecom industry. In view of this, Customer Relationship Management (CRM) has become a critical and comprehensive strategy for managing and interacting with customers to improve customer retention and also to attract potential customers.

Guo-en et al. [1] estimated that the average churn rate for the mobile telecommunication is 2.2% per month. Telecom companies spend hugely in acquiring new customers every year and so not only does a company loose future revenue when a customer switch provider (churn) but also the resources spent in acquiring the customer. To combat this challenge, researchers and industries had to advent to a more sophisticated data mining techniques rather than the traditional methods. The knowledge discovery in data (KDD) technology has made it possible to derive more brilliant and advanced knowledge from large data collections. KDD obtain knowledge by extracting patterns from data. The patterns generated could be used to help companies foresee the likelihood of a customer to churn and therefore develop better retention measures [2, 3].

In Customer Churn Predictions (CCP), customers are categorized into two sets of classification behaviors known as churn and Non-churn. A customer is categorized as a churner when they switch from one service provider to the other. Non-churn customers contrarily are loyal customers who do not switch provider and may buy new product or services (appetency), and/or buy upgrades or add-ons (up-selling). Appetency and up-selling are important strategies in customer retention and when done correctly can benefit telecom companies in the following: (i) Generate more revenue, (ii) bring in new customers, (iii) help retain customers longer, (iv) get customers to spend more, (v) give customers a fair value of what they deserve.

However, appetency and up-selling have not widely been researched in the telecommunication sector. To bridge this research gaps, this study aims to develop an accurate and comprehensible prediction scheme for appetency and up-selling predictions using machine learning algorithm via support vector machine (SVM). SVM algorithm is implemented in this study because it has been proven to be an effective classification approach for CCP [4] compared with Artificial Neural Network

(ANN), decision tree, logic regression and Na ïve Bayesian classifiers.

II. Related Works

Since customer churn is an important issue, several authors have been conducting investigations and proposing methods for churn prediction. Ammar et al. [5] presented a meta-heuristic based CCP approach using a hybridized form of firefly algorithm as the classifier. This approach compares every firefly with every other firefly to identify which one has the highest light intensity. The hybridized firefly algorithm provides effective and faster results. Bloemer et al. [6] suggested that a greater degree of customer satisfaction enhance the general performance of a company in the highly competitive telecom industry. Customer satisfaction has in recent years been the focus of most companies thereby giving birth to various machine learning techniques being applied for CCP. Amin et al. [2] proposed a rough set theory (RST) CCP method utilizing Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA), and the LEM2 Algorithm. The results indicated RST based on GA as the overall best performed method for knowledge extraction from publicly available telecom dataset. Caigny et al. [7] proposed a new hybrid algorithm, the logit leaf model (LLM) for better data classification. The proposed algorithm consists of two stages: a segmental phase where customer segments are identified using decision rules and a prediction phase where a model is created for every leaf of the tree. In this approach, LLM outperformed its building blocks i.e. logistic regression and decision tree with a significant predictive score.

SVM is a supervised learning method used for regression and classification analysis. It was developed at AT & T Bell laboratory by Vladimir Vapnik and his co-worker [8]. It can be applied to pattern recognition, static function approximation and regression. SVM minimizes the upper bound on the actual risk based on the Structural Risk Management (SRM) principles in contrast to other classifiers which seek to minimize the empirical risk [9]. From a survey carried out by Umayaparvathi et al. [10] on the most frequently used CCP technique, they found SVM to out-perform neural network and decision tree. Kamya et al. also found SVM to be the most used CCP technique mostly due to its ability to be applied to both regression and classification. Niccolo et al. [11] proposed a CCP model using SVM based Area Under the Curve (AUC) method (SVMauc). The proposed method employed SVMauc parameter-selection for churn prediction. The results indicated that, SVMauc optimizes the generalization performance and the parameter-selection improved the prediction performance. Guo-en et al. [1] proposed SVM approach aimed at minimizing the structural risk to improve prediction accuracy. The approach aims at detecting the infrastructure risks and determining the relations between them and customer churn. This approach enjoys high accuracy rate even with abundant attributes and nonlinear data.

III. Methodology

In this section, the appetency and upselling classification scheme based on SVM is presented. We proceed to explain the theory behind SVM and the various SVM kernel functions that are adopted in this work. Ultimately, the basis of SVM posterior probability and its application in the proposed prediction scheme are detailed.

A. Support vector machine (SVM)

Consider a classification task, in which data is separated into training and testing sets. Each instance in the training set contains one class label and several corresponding features or attributes. In SVM, the goal is to use the training data to produce a model which can predict the class label of each independent instance in the testing data based on their features or attributes. Let $(x_i, y_i), i = 1, \ldots, l$ be a training data set, where $x_i \in R^n$ and $y_i \in \{1, -1\}^l$. n is the number of attributes of each input x_i, y_i is the class label of input x_i and l is the number of training points [12]. Supposing the data is linearly separable, a hyperplane of the form described in (1) can be used to separate the two classes.

$$w.x + b = 0 \tag{1}$$

where w is a weight vector, x is input vector and b is bias [13]. The support vectors are the data points in the training set closest to the hyperplane. SVM finds the optimal hyperplane such that the separation between that hyperplane and the support vectors is maximized. Eliminating any or all of the support vectors would change the position of the optimal separating hyperplane, since they are the critical elements of the training set. In order to apply SVM to classify each instance in the testing data, the solution of the following optimization problem is required:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

Subject to

$$y_i \left(w^T \varphi(x_i) + b \right) \geq 1 - \xi_i \tag{2}$$

where $\xi_i \geq 0$. φ is a function which maps the training vectors x_i into a higher dimensional space such that SVM finds a linear separating hyper plane with the maximal margin between the classes in this space. $C > 0$ is a regularization parameter of the error term [14]. ξ_i is a positive slack variable which is applied to enable SVM handle data that is not completely linearly separable.

B. Kernel Functions

Instead of using the original input attributes x_i in SVM, some features $\varphi(x_i)$ may rather be applied to get SVMs to learn in the high dimensional feature space [15]. Considering a feature mapping φ , the corresponding Kernel can be defined mathematically as

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \tag{3}$$

$K(x_i, x_j)$ could be easily solved by determining $\varphi(x_i)$ and $\varphi(x_j)$ computing their inner product without needing to identify or represent the vectors $\varphi(x_i)$ explicitly. $K(x_i, x_j)$ is a measure of the similarity between $\varphi(x_i)$ and $\varphi(x_i)$, and consequently between the two independent attributes x_i and x_j. SVM makes predictions using a linear combination of kernel basis functions. The kernel defined by the function in (3) is known as the linear kernel and can work perfectly well for linearly separable data. On the other hand, non-linear kernels are required when data is not fully linearly separable. Therefore, we use a non-linear SVM kernel, the polynomial kernel, in addition to the linear kernel for classification purposes in this paper. The polynomial kernel is one of the quintessential non-linear kernels and can be defined by

$$K(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^d, \gamma > 0 \tag{4}$$

where γ, r and d are the parameters defining the kernel's behavior [16].

C. Posterior probability

Posterior probability is a theory in Bayesian statistics used to determine the probability of an event given the knowledge of occurrence of other events that bear on it. It can be defined as the conditional probability of a random event or an uncertain proposition after considering key background information on randomly observed data [17]. Contextually, the term 'posterior' relates to a hypothesis of what can be ascertained through an understanding of how certain things occur, taking into account the relevant evidence. Posterior probability may be applied in classification to indicate the uncertainty of placing an observation in a particular class.

In machine learning, transforming class membership values into posterior probabilities allow for comparison and post-processing [18]. For SVM, the posterior probability is a function of the score $P(s)$ that an instance j is in class $y_j = \{-1,1\}$. In the case of linearly separable data sets, the posterior probability is evaluated according to the step function (5).

$$P(s_j) = \left\{ \begin{array}{ll} 0, & S_j < \max_{y_i=-1} S_i \\ \pi, & \max_{y_i=-1} S_i \leq S_j \leq \max_{y_i=+1} S_i \\ 1, & S_j > \max_{y_i=+1} S_i \end{array} \right\} \tag{5}$$

where s_j is the score of the instance $j; +1$ and ? denote the positive and negative classes, respectively; π is the prior probability that an instance is in the positive class. The following sigmoid function computes the posterior probabilities for linearly inseparable data classes:

$$P(s_j) = \frac{1}{1 + exp(As_j + B)} \tag{6}$$

where the parameters A and B are the slope and intercept parameters, respectively [19].

D. The Prediction Scheme

In customer churn prediction, posterior probability using SVM is an indication of the probability of a customer to churn or remain with a service provider such as a mobile network operator. For customers who do not churn, they may want to buy new products or an upgrade. In other words, for a randomly selected customer, there could be a conditional probability of buying a new product or an upgrade only if he remains with the network provider. In this paper, a data set of customers' attributes and class labels (churn or non-churn) is used to train SVM algorithms so that it can make customer churn predictions based on randomly distributed testing data. The class membership scores that SVM evaluates to classify customers are subsequently transformed into posterior probabilities to make appetency and up-selling predictions. Our proposed prediction scheme leverages the posterior probability of belonging to a non-churn class to determine the level of a customer satisfaction. We postulate that the posterior probability is an indication of the level of customer's satisfaction with a mobile network provider's services. Customer satisfaction level is derived from a satisfaction score, which is calculated from the posterior probability of belonging to the non-churn class. We obtain the customer satisfaction score c of a customer j as $c_j = \beta * P(s_j)$, where $P(s_j)$ is posterior probability of customer j being in the non-churn class based on its membership score s and β=10 is a score normalization factor. Based on a customer satisfaction level, we can make future inferences about his appetency and up-selling behavior as shown in Table 1.

Table 1. Customer satisfaction scores and attributions

Customer Satisfaction Score	Customer Satisfaction Level	Potential Customer Attribute
8-10	Very Satisfied	• Remain • Upgrade • Buy new product
6-7.99	Satisfied	• Remain • Upgrade
4-5.99	Okay	• Remain
2-3.99	Not Satisfied	Switch but may return
0-1.99	Very Dissatisfied	Switch and may never return

IV. Empirical Analysis

A. Data preparation and feature selection

Acquiring actual dataset from telecom industries can be a great challenge due to customer privacy. Nonetheless, there are publicly available datasets for data analysis. Access to the dataset used in this study can be found at the University of California, Irvine, telecom dataset, UCI repository [20]. With 3333 instances consisting of 2850 (85.5%) non-churn (NC) and 483 (14.49%) churn (C) customers, this dataset is considered ideal for modeling. A series of experiments based on the proposed prediction framework is conducted out using MATLAB toolkit. Fig. 1 is a visual presentation of the proposed framework.

Table 2. Attribute description

Attribute	Description
account length	No. of days a customer has been using the service
Intl_Plan	If a customer has international plan or not
VMail_Plan	If a customer has voice mail plan or not
VMail_Msg	Number of voice mail messages
Day_Mins	Daytime minutes used by the customer
Day_Calls	Daytime calls used by the customer
Eve_Mins	Evening time minutes used by the customer
Eve_ Calls	Evening time calls used by the customer
Night_Mins	Night time minutes used by the customer
Night_calls	Night time calls used by the customer
Intl_Mins	Minutes of calls made whiles abroad
Intl_Calls	Calls made whiles abroad
CustSer_calls	Number of calls made to customer care center
Churn?	1 for churners and 0 for non-churners

To reduce computational cost, feature selection has become an important process in knowledge discovery [21, 22]. There are 21 attributes in the dataset used in this study. However, not all attributes in a dataset are suitable for modeling [23]. Some attributes are unique and contain no predictive value, therefore they cannot be used in modeling. In this dataset, the “state, area code, phone number” presents customer information, and “the four charge attributes” have been eliminated since they do not contain relevant information for prediction [4]. Categorical values are normalized by converting ‘yes’ or ‘no’ and ‘true’ or ‘false’ into 1s and 0s. 3. Oversampling was performed on the training set by duplicating the minority class (churn) to obtain almost equal number of minority to majority class. Table 2 describes the influential attributes used in the modeling process.

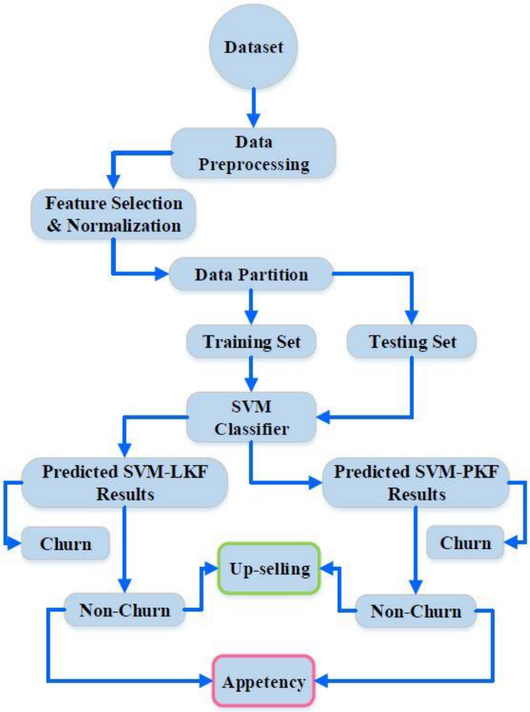

Fig. 1. Visualization of proposed prediction framework.

B. Data Partition

In other to evaluate the performance of a trained prediction model, a different set of data other than the one it was trained on is required for validating. This new set of data is the validation set. Validation ensures the model remembers the instances it was trained on and to perform well on unseen new instances [24]. In this paper, we presented a random sampling sizes of (70 – 80) % training set to (30 - 20) % test set for finding how they impact the classifier’s decision. Table 3 describes the training and test datasets after partitioning.

Table 3. Distribution of training and test set

Training set (%)	Observations	Test set (%)	Observations
80	2667	20	666
77	2567	23	766
75	2500	25	833
72	2400	28	933
70	2333	30	1000

V. Results and Discussions

This section explores the performance of the proposed study and evaluates the results of the different kernel functions through standard evaluation measures. In subsection A, the churn predictions accuracy results of both models are discussed. Subsection B presents the predicted results of the number of likely up-selling and appetency customers. The least minimum predicted errors of both kernels in subsection C.

A. *Churn Accuracy Prediction*

In this section, the accuracy results of both kernels in terms of correctly predicted churn and non-churn customers are evaluated. It is observed that both proposed models exhibit different performance using different sample sizes. The linear kernel however, has the lowest accuracy of 83.203% using a sample size of 77:23%. Fig. 2 demonstrate that polynomial kernel function obtained the best prediction accuracy of 91.67% using a sample size of 77:23% and an overall accuracy of (above 90%) as compared to linear kernel function which predicted an accuracy of 84.19% using a sample size of 75:25% and an overall accuracy of (above 83%).

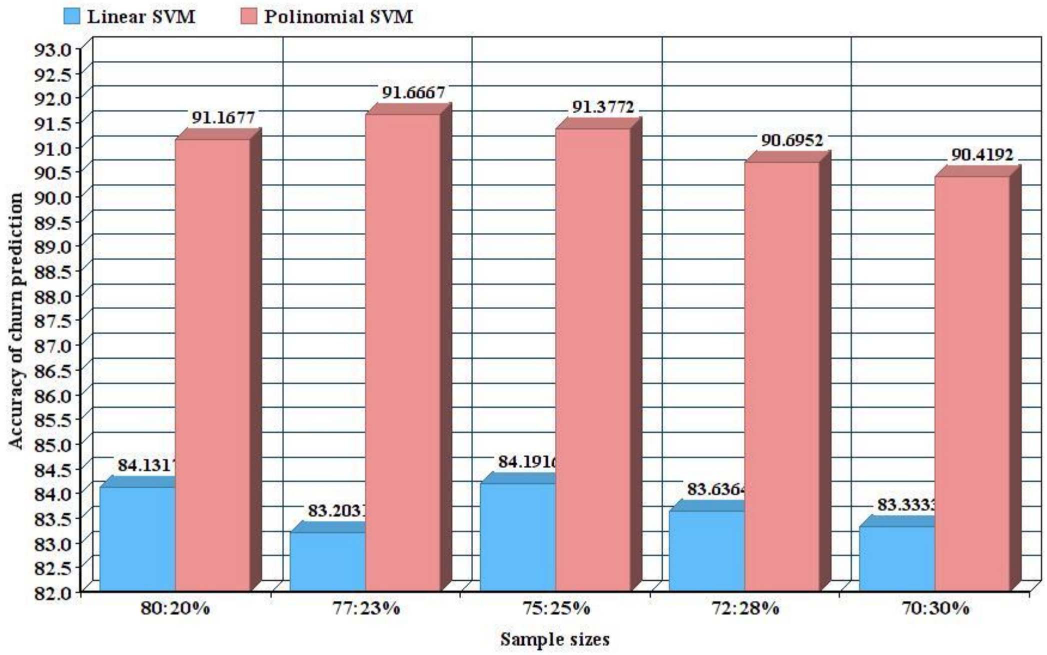

Fig.2. Accuracy of churn prediction

B. *Number of likely up-sell and appetency customers*

In general, the number of predicted appetency and up-sell customers increases when the sample size is reduced because there is an increase in non-churn predictions. However, for linear kernel, a change in sample size from (77:23 to 75:25) % slightly reduces the number of predicted appetency customers. Using a sample size of 70:30, the linear kernel predicted the maximum number of upselling (988 customers) and appetency (817 customers) as shown in tables 4 and 5.

Table 4. No. of likely up-sell customers

Sample sizes (%)	Linear	Polynomial
80:20	627	572
77:23	767	654
75:25	786	715
72:28	925	803
70:30	988	863

Table 5. No. of likely appetency customers

Sample sizes (%)	Linear	Polynomial
80:20	510	522
77:23	690	598
75:25	677	651
72:28	764	735
70:30	817	793

C. Minimum error (ME)

This section discusses the results of the ME of the prediction models. These errors are generated when a churner is rather predicted as a non-churner. A wrong non-churn prediction will lead to an error in appetency and up-selling predictions. Table 6 reflects the comparison of predicted errors of our proposed models in up-selling and appetency using different sample sizes. It can be seen that the polynomial kernel predicted the least ME value in up-selling of 0.0587 using a sample size of 75:25% compared with the linear kernel with least ME value of 0.1388 using a sample size of 80:20% in up-selling. Also in appetency, polynomial kernel predicted the least ME value of 0.0364 using a sample size of 80:20%. The linear kernel on, the other hand, predicted a least ME value of 0.0765 using a sample size of 80:20%. In general, polynomial kernel function predicted the least ME across all sample size.

Table 6. Me of linear and polynomial kernel functions

	Up-selling		Appetency	
Sample sizes	Linear	Polynomial	Linear	Polynomial
80:20	0.1388	0.0594	0.0765	0.0364
77:23	0.1669	0.0596	0.142	0.0385
75:25	0.1412	0.0587	0.099	0.0369
72:28	0.16	0.0648	0.127	0.0422
70:30	0.1609	0.0718	0.1248	0.0492

VI. Conclusion

This paper presents a novel up-selling and appetency prediction scheme using SVM algorithm with two different kernel functions. After evaluation of the results, it was observed that polynomial kernel function had the highest accuracies (above 91%) for the first three sample size (i.e. 80:20, 77:23, 75:25) % in predicting both churners and non-churners compared with the linear kernel function. The proposed scheme speculates using the posterior probability of customers whether a non-churner will buy a new product or buy an upgrade. Again the polynomial kernel function predicted the least minimum error and therefore is considered to be the best model for appetency and up-selling predictions in telecommunications. Also using different sampling sizes revealed that a range of sample sizes equally have an effective prediction performance on the model. This will be beneficial to both researchers and mobile companies to incorporate different sample sizes rather than focusing on the sample size that obtained the highest accuracy. Future studies could be to investigate with other models and compare to our results for statistical evaluation. Furthermore, more data on customer appetency and up-selling behavior will be examined.

References

[1] Xia, G.-e. and W.-d. Jin, Model of Customer Churn Prediction on Support Vector Machine. Systems Engineering - Theory & Practice, 2008. 28(1): p. 71-77.

[2] Amin, A., et al., Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 2017. 237: p. 242-254.

[3] Rodan, A., et al., A Support Vector Machine Approach for Churn Prediction in Telecom Industry. Vol. 17. 2014.

[4] Ionut Brandusoiu, G.T., Churn Prediction in the Telecommunications Sector using Support Vector machines. Annals of the University of Oradea, 2013. Volume xxii (xii), 2013/1.

[5] Ahmed, A. and D. Maheswari Linen, A review and analysis of churn prediction methods for customer retention in telecom industries. 2017. 1-7.

[6] Bloemer, J., K. de Ruyter, and P. Peeters, Investigating drivers of bank loyalty: the complex relationship between image, service quality and satisfaction. 1998. 16(7): p. 276-286.

[7] De Caigny, A., K. Coussement, and K. De Bock, A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. Vol. 269. 2018.

[8] Vapnik, V.N., The nature of statistical learning theory. 1995: Springer-Verlag. 188.

[9] Coussement, K. and D. Van den Poel, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 2008. 34(1): p. 313-327.

[10] V. Umayaparvathi, K.I., A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. International Research Journal of Engineering and Technology (IRJET), 2016. 03(04).

[11] Gordini, N. and V. Veglio, Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Industrial Marketing Management, 2017. 62: p. 100-107.

[12] Fletcher, T., Support Vector Machines Explained. 2009.

[13] R. Berwick, V.I., An Idiot’s guide to Support vector machines (SVMs) 2003.

[14] Ng, A., CS229 Lecture notes. 2000.

[15] Noble, W.S., What is a support vector machine? Nature Biotechnology, 2006. 24: p. 1565.

[16] Hsu, C., C. Chang, and C. Lin, A practical guide to support vector classification. Vol. 101. 2008. 1396-1400.

[17] Duan, K., et al., Multi-Category Classification by Soft-Max Combination of Binary Classifiers. 2003. 125-134.

[18] Qing, T., et al., Posterior probability support vector Machines for unbalanced data. IEEE Transactions on Neural Networks, 2005. 16(6): p. 1561-1573.

[19] Platt, J., Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Vol. 10. 2000.

[20] Becks, D. Churn in Telecom's dataset. 2018 [cited 2019 25]; Available from: https://www.kaggle.com/becksddf/churn-in-telecoms-dataset/version/1.

[21] Stojanović, M.B., et al., A methodology for training set instance selection using mutual information in time series prediction. Neurocomputing, 2014. 141: p. 236-245.

[22] Malik, Z.K., A. Hussain, and J. Wu, An online generalized eigenvalue version of Laplacian Eigenmaps for visual big data. Neurocomputing, 2016. 173: p. 127-136.

[23] Dr. M. Balasubramanian , M.S., Churn Prediction in Mobile Telecom System using Data Mining Techniques International Journal Of Scientific And Research Publilcations, 2014. 4(4).

[24] Bellazzi, R. and B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform, 2008. 77(2): p. 81-97.

Authors' Profiles

Zhou Lian-Ying received her MSc. From Jiangsu University of Science and Technology and her Ph.D. from Nanjing University of Science and Technology in 1997 and 2008 respectively. She is currently a professor at the school of computer science and communication engineering, Jiangsu University, Zhenjiang, China. She has published more than 30 articles in international and national journals. Her research areas include but not limited to Computer Networks, Electronic commerce, Intrusion Detection technology, Internet of Things, and Bid Data Analytics.

Daniel M. Amoh completed his B.Sc. degree in Computer Engineering from Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana in the ccyear 2015. He is presently a Master's student at the school of computer science and communication engineering, Jiangsu University, Zhenjiang, China. His area of research includes Computer Networks, Big Data Analytics and Machine Learning.

Louis K. Boateng has completed his BSc. in Computer Engineering from Kwame Nkrumah University of Science and Technology (KNUST) and was awarded a degree in 2012. He is currently undertaking his MSc. in Information and Communication Engineering at the school of computer science and communication engineering, Jiangsu University, Zhenjiang, China. He majors in Big Data Analytics and Machine Learning.

Andrews A. Okine received his B.Sc. degree in Telecommunication Engineering from Kwame Nkrumah University of Science and Technology (KNUST) Kumasi, Ghana, in June 2014. Since September 2016, he has been working towards his master's degree at the school of computer science and communication engineering, Jiangsu University, Zhenjiang, China. His research interests include Optical Wireless Communications, Heterogeneous Networks and Machine Learning for Wireless Communication.

How to cite this paper: Lian-Ying Zhou, Daniel M. Amoh, Louis K. Boateng, Andrews A. Okine, "Combined Appetency and Upselling Prediction Scheme in Telecommunication Sector Using Support Vector Machines", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.6, pp. 1-7, 2019.DOI: 10.5815/ijmecs.2019.06.01