Building Predictive Models to classify the right segment of customers by predicting click and open probabilities of an email campaign

Ashok Shetty
PGDM in Business Analytics,
REVA University, Bengaluru
ashokshetty.ba02@reva.edu.in

Lakshmi D
MBA in Business Analytics,
REVA University, Bengaluru
Lakshmid.BA01@reva.edu.in

Abstract – Email marketing is the extraordinarily powerful virtual marketing approach of sending emails to prospects and potential customers. Effective e-mail marketing converts possibilities into customers and turns one-time buyers into dependable, raving lovers. Despite the rise of social media and unsolicited spam, email still remains the most effective way to nurture the customers. Companies are spending a lot of time in writing that perfect email, laboring over each word, innovative subject lines, senders name and catchy layouts on multiple devices to get them best in-industry click rates & view rates. Objective of this study is to analyze if we can optimize our email marketing campaigns using different analytical techniques for predicting the click probability of email campaigns.

For the purpose of our study, we have used the data sets from the "Lord of the Machines" Data Science hackathon from Analytics Vidhya. The datasets used in this hackathon contained real-life cases, where more than 1 million emails/news letters sent by Analytics Vidhya for promotional and campaigning purposes were included. The data was shared after masking/removing some personal and sensitive information. This enabled us to work on a real-life problem, currently being encountered by the email marketing teams in various domains. The solution was to come up with a predictive model that lets us classify between the users based on their click probabilities and thereby, selecting the right segment of customers for the different types of campaigns. To come up with the best solution for this challenge we have used various supervised machine learning algorithms such as Naïve Bayes, Logit, SVM, Random Forest etc. (which also involved a bit of text mining).

Keywords: Email campaign, real- life problem, Naïve Bayes, SVM, Random Forest

I. INTRODUCTION

Objective desires – is the main approach for all excellent marketing starts and electronic marketing's no distinct. For making an effective email marketing campaign it is very much necessary to understand the typical goals e. g. welcoming new subscribers, boosting engagement, nurturing existing subscribers, re-engaging subscribers, segmenting subscribers etc. Sending an email is just the first step in achieving email marketing success, to really nail it, campaigners got to collect data to improve future campaigns. That means testing everything: design and layout, email marketing copy, subject lines and calls to action and also considering the testing emails with different segments and experimenting with email send times, too.

Email campaigners also need to consult with email analytics from the service provider relating to opens, clicks, unsubscribes and forwards. This will enable them to figure out what's working and what's not with email campaigning [1]. Another major issue which affects email deliverability is 'sender reputation'. Hence campaigners should use 'sender score' to see if there are any red flags which are stopping emails from subscribers' inboxes. Finally managing email subscriber list by attempting to re-engage inactive subscribers or removing them if the attempts fail. It's better for email marketing open and click rates to have fewer active subscribers than large numbers of inactive ones.

II. THE NEW CHALLENGE

Let's say a Digital Marketing team would like to explore the possibilities of applying Machine Learning models of millions of past records which were sent out via emails and which contain data on products, customers, and email details including an indicator which says whether or not an email has been opened (or links inside clicked on). In this case applying some Regression models can definitely help for meeting the objective, however the reality, little of regression analysis might churned out any useful actionable points to integrate into decision making. For the marketing people, they might just wanted to 'tweak'

the campaigns so that email engagement with customers will be maximized. By telling them that a male customer has higher rate to open an email than a female customer is not going to be very helpful [2].

And this is where the Machine Learning has come into picture. Even though no algorithm can 'learn' to draft a perfect email which maximizes the chance of opening for every customer, however just to know a probability or a number which tells how likely a customer is going to open the very next mail which can actually help the campaigners to properly differentiate the customers as well as target them separately with different marketing channels [2].

III. LITERATURE REVIEW

Over the beyond decades on line advertising has grown to $70 billion industry international yearly. Despite of the impressive growth, marketing campaigners faces many challenges to measure the effective campaigns [3]. In the present era of digital world, email campaigning is one of the most effective and popular marketing media. However finding the exact rate of clicking or opening the emails by the customers is the most critical factor of evaluating the effectiveness of email campaigning.

The advantages of Email marketing have been diagnosed by way of many unique authors. Forrester (Niall 2000) describes electronic mail campaigning as one of the most simplest and effective online tool because of its excessive reaction rate and was expecting email marketing to be worth 5 billion US dollars by 2004 [4]. In one of the research paper, Di lanni looked at how an e-enterprise 'virtual' agency, one generally reliant on the Internet for all interaction and communication with its audiences, can broaden and setups sophisticated marketing campaigns the usage of the Internet as the most important medium [5].

J Stern In their record Opt-in Email Gets Personal Forrester Research (www. forrester.com) stated opt-in email "will spread like wildfire." They consider using opt-in email for campaigning "will explode" because businesses may be lured by means of excessive rates, low prices, and the convenience with which any firm, huge or small, can get started out [6]. Luo, Nadanasabapathy, Zincir-Heywood, Gallant in their studies paper presented stories using a getting to know model on predicting the "opens" and "unopens" of segmented marketing emails. The version became primarily based at the features extracted from the emails and email recipients profiles. To attain this, they employed and evaluated unique classifiers and one of a kind data sets using distinctive characteristic units. Final consequences confirmed that it is viable to predict the rate for a targeted marketing e mail to be opened or not with about 78 % F1-measure [7].

Few authors (Rettie, Grandcolas, Deakins) explained about text message advertising which is a form of telemarketing and which shares the features with email marketing [8]. In one of the research paper Pavlov and Melville explained about chance of being overrun with the aid of unwanted industrial e mail (also referred to as junk mail). In order to recognize the underlying dynamics of the spam industry and to take a look at alternative mitigation techniques, they expand a device dynamic model which famous that the gadget conforms to the boundaries-to-boom normal shape. Simulations suggest that filtering might also have the unintended outcome of growing the worldwide quantity of spam. The unexpected increase comes about due to the fact better filters can sincerely help spammers by abating records deficit [9].

Few researchers cautioned a unique approach named Urban POI-Mine (UPOI-Mine) that integrates region-based totally social networks (LBSNs) for recommending users urban POIs based at the person choices and region houses concurrently. The idea of UPOI-Mine is to construct a regression –tree –based model in normalized take a look at in space which helps the prediction of POI related to each different alternatives. Based at the LBSN information, diverse researchers can perceive one of a kind functions of places consisting of (1) Social factor (2) Individual alternatives and (3) POI reputation for model building. As in keeping with the take a look at, the said study turned into the primary work on city POI advice which considers social aspect, man or woman preference and POI popularity in LBSN facts. By the use of this real set of information from Gowalla, the proposed UPOI-Mine confirmed to supply awesome performance [10].

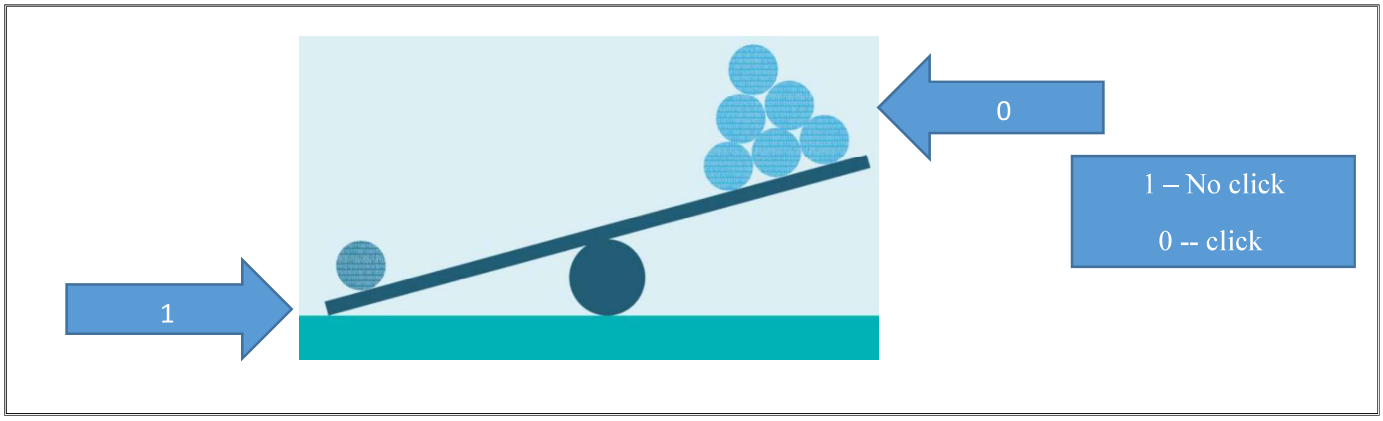
DATA DESCRIPTION

The dataset used for analysis from the "Lord of the Machines" Data Science hackathon from Analytics Vidhya. The datasets includes real life cases which contained 1 million email/news letters sent by Analytics Vidhya form promotional and campaigning purposes. The solution was to come up with a predictive model that lets us classify between the users based on their click probabilities and thereby, selecting the right segment of customers for the different types of campaigns.

The dataset contains 3 sets of data – Train data and Test data which contains approximately 1.3 million observations with 6 variables e.g. send_date, is_open, is_click, send_time, time_of_day, time_of_month, communication_type etc. as well as campaign data count of 53 which includes text information like subject line, body of the email, communication type, URL links etc. To come up with the best solution for this challenge we have used various supervised machine learning algorithms such as Naïve Bayes, Logit, SVM, Random Forest etc and also little bit of text mining for campaign data.

IV. DATA PREPARATION

The dataset available for analysis had approximately 1.3 million observation with 6 feature variables, no missing values. However data sets had huge class imbalance problem where approx. 99% were for no click and only 1% for clicking probability; to resolve the issue we used SMOTE technique.



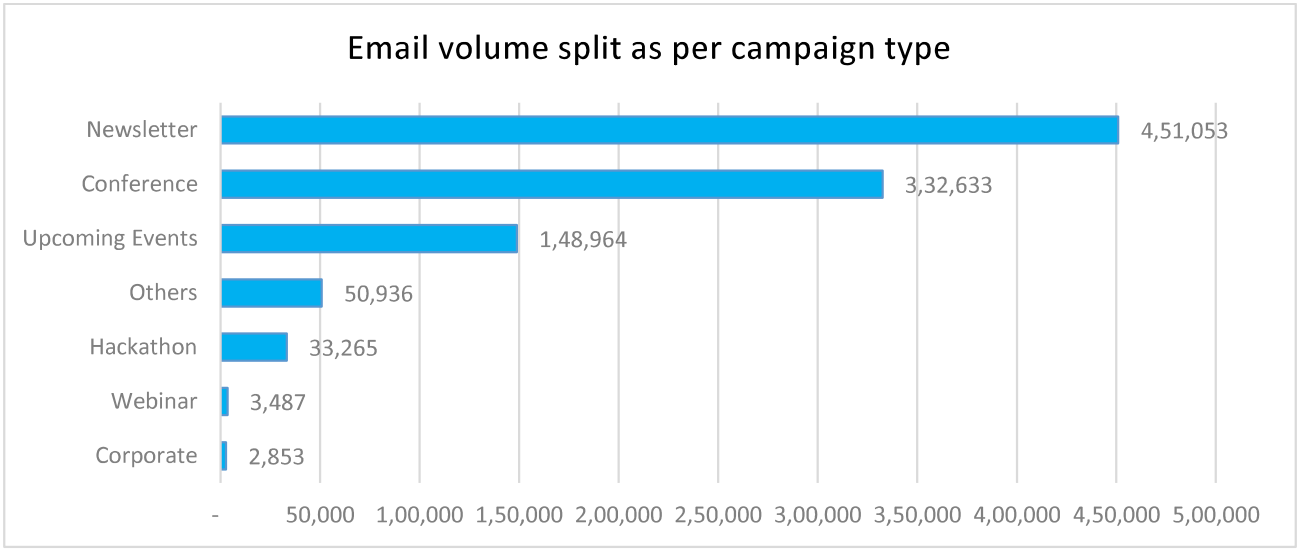
To do the data cleaning and casting which follows correcting date format as lot of 'dots' were existed in the data. We also derived new variables (to get the categorical variables) as per below-

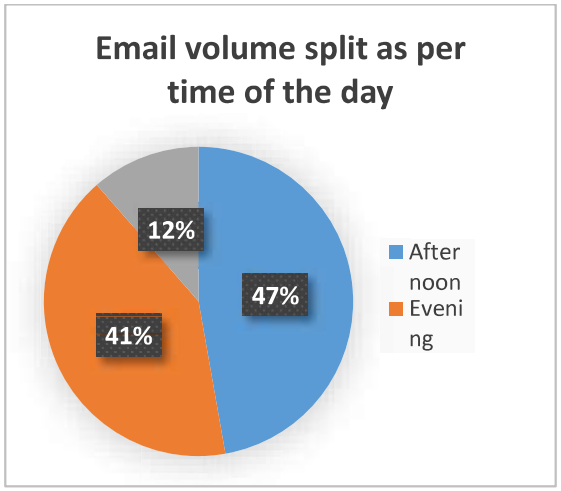
- Day of the week - when the email was sent (Mon, Tue, Wed etc.)
- Time of the day - when the email was sent (Morning, Afternoon and Evening)
- Time of the month - when the email was sent (First half or Second half)
- Hit rate – Number of emails read / total email sent for a specific campaign

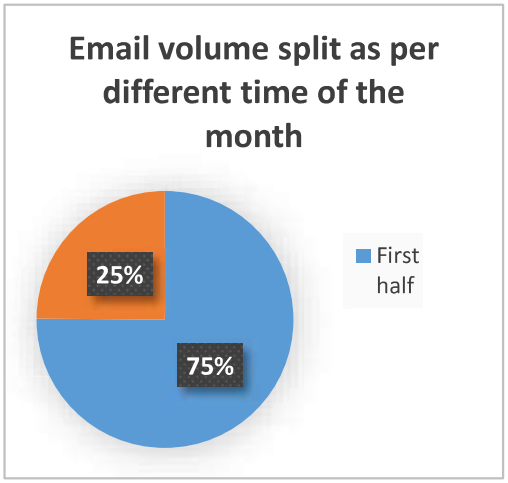
Hackathon also provided campaign dataset which contained text messages of 53 rows. Hence to convert this unstructured data to structured format we followed few steps as below-

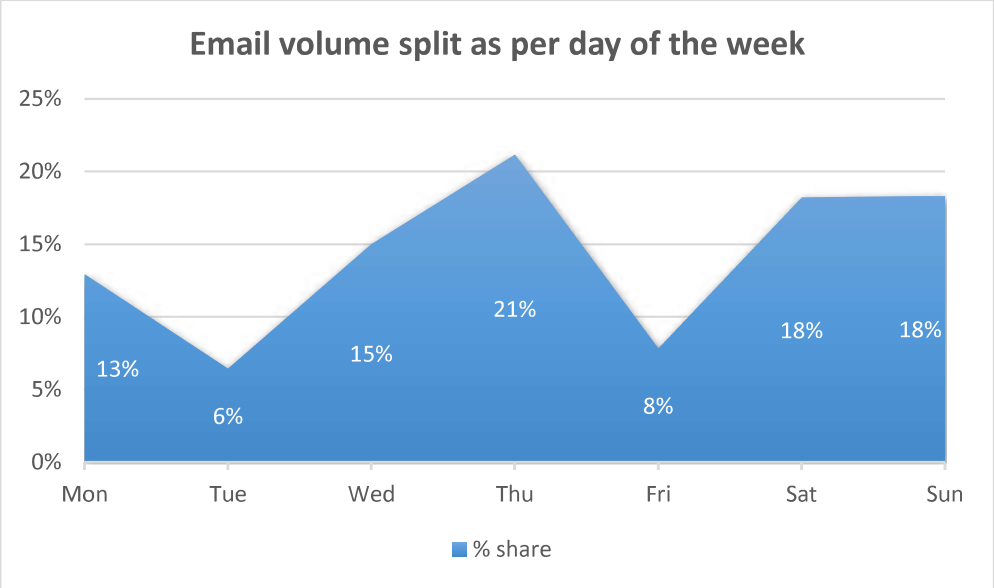
- Created a corpus of the email text to concatenate the email body and subject line
- For data cleaning:
  - Converted the text data to lower case
  - Removed the punctuations
  - Getting rid of numbers, stop words, white spaces
  - Stemming words
- Created a Document Term Matrix(DTM) which enable us to get the unique words
- Finally merged DTM and variables so that we can get the maximum number of words which are clicked
- Fitting models and comparing accuracy

Graphical representation of Data









V. METHODOLOGY

The objective of the study is to build different models to predict the click rate on the emails so that an effective email campaigning could be possible by the marketing team which can save time and cost. To build different models various processes involved which includes data cleaning, missing data check, reformatting days/months etc. to a readable format by R, converting unstructured data to a structured format etc. Also as the dependent variable is having unbalanced data hence we used SMOTE to balance the same.

The available dataset already contained train data, test data and campaign data; so we used directly the respective set of data files for building the predictive models and validated the same.

The following different predictive models were built-

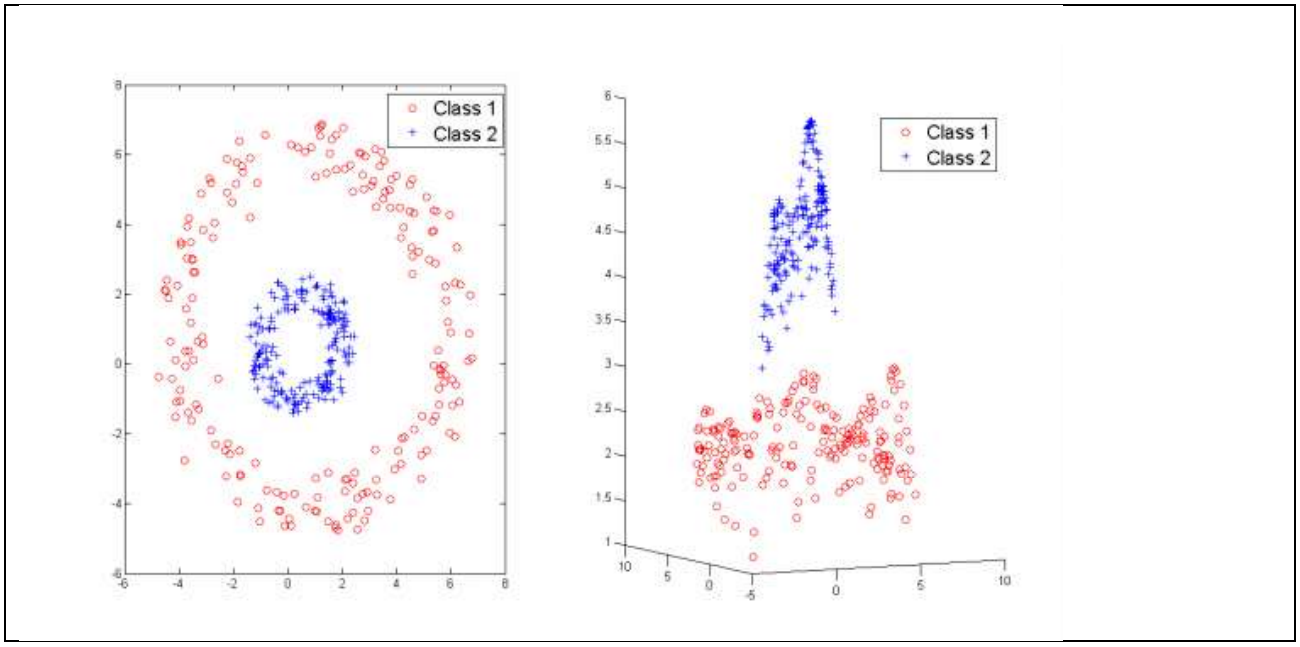
- Naïve Bayes
- Logistics Regression
- Random Forest
- SVM
- XGBoosting

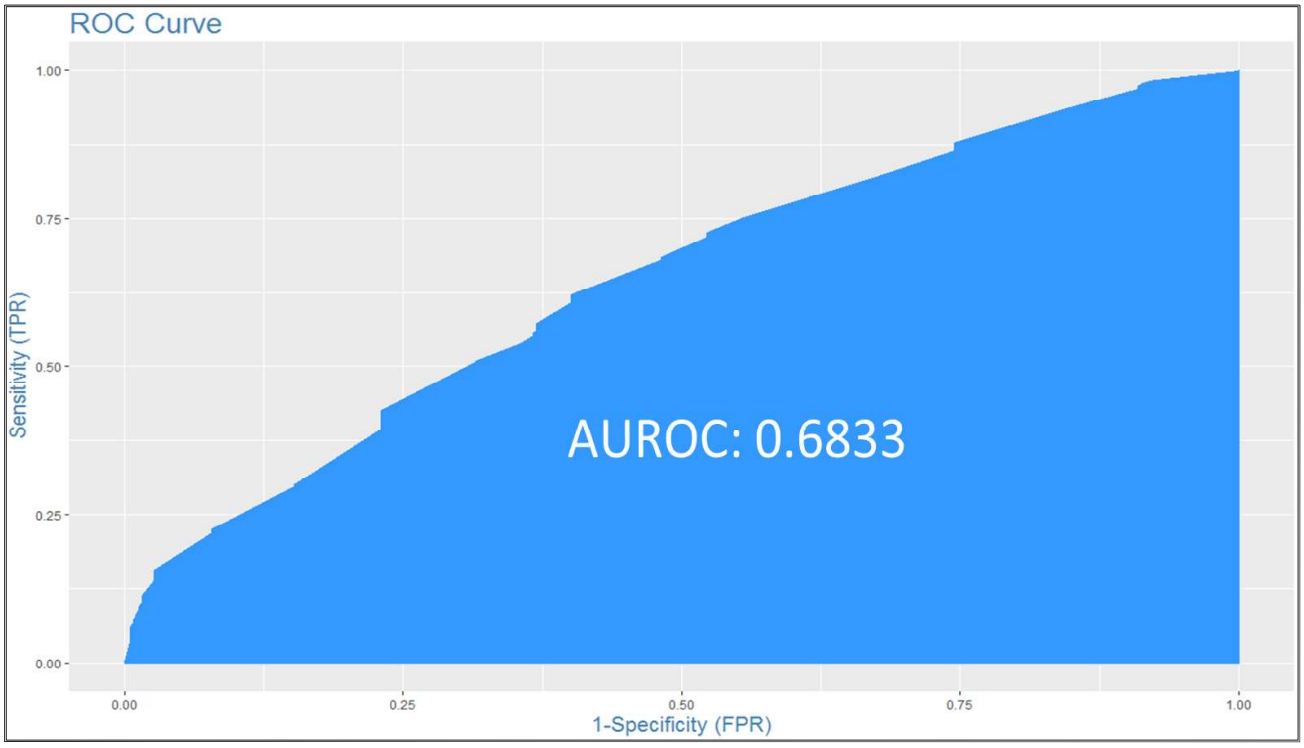
The best classifier after testing several different types is - SVM (RBF kernel)

1. The accuracy of the model (as requested by the organizing committee) is: 0.68 (AUC) with SVM RBF kernel.

2. The parameters of it are: gamma=0.5, C=3

The intuition of SVM - RBF kernel:





VI. FINDINGS/DISCUSSION

While comparing classification models, SVM(RBF Kernal) could give the best accuracy of the model which is 68% where Gamma=0.5 and C=3 (we tested with all 9 different kernals- rbfdot, polydaot, vanilladot, tanhdot, hyperbolic, laplacedot, besseldot, anovadot, splinedot, stringdot). Whereas the other

models were not able to improve the model accuracy even though we used few ensambling techniques.

Below are the results of different predictive models-

Algorithm	Accuracy (AUC)
Naïve Bayes	0.63
Logistic Regression	0.63
Random Forest	0.65
SVM	**0.68**
XGBoosting	0.66

CONCLUSION/IMPLICATIONS

Based on the model accuracy it is evident that SVM (RBF kernel) predictive model can better perform in comparison to Naïve Bayes, Logistics Regression or other classification techniques. Therefore to minimize the time and cost by sending effective emails marketing campaigning team can utilize SVM technique to get the better clicking probability rate. However in future we would like to test couple of more algorithms such as RNN and LSTM on the same dataset to see whether they can perform better than SVM.

REFERENCES

1. https://optinmonster.com/how-to-run-a-successful-email-marketing-campaign/
2. https://weiminwang.blog/2016/12/22/apply-machine-learning-to-email-campaign/
3. Farahat, A., Shanahan, J.: Econometric analysis and digital marketing: how tomeasure the effectiveness of an ad. In: ACM WSDM (2013)
4. Rettie, R. (2002). Email marketing: success factors.
5. Di Ianni, A. (2000). The e-business enterprise and the 'Web-first'principle of e-marketing. *Interactive Marketing*, *2*(2), 158-170.
6. Sterne, J., & Priore, A. (2000). *Email marketing: using email to reach your target audience and build customer relationships*. John Wiley & Sons, Inc..
7. Luo, X., Nadanasabapathy, R., Zincir-Heywood, A. N., Gallant, K., & Peduruge, J. (2015, October). Predictive Analysis on Tracking Emails for Targeted Marketing. In *International Conference on Discovery Science* (pp. 116-130). Springer, Cham
8. Rettie, R., Grandcolas, U., & Deakins, B. (2004). Text message advertising: dramatic effect on purchase intentions.
9. Pavlov, O. V., Melville, N., & Plice, R. K. (2008). Toward a sustainable email marketing infrastructure. *Journal of Business Research*, *61*(11), 1191-1199.
10. Erheng Zhong , Ben Tan , Kaixiang Mo , Qiang Yang, User demographics prediction based on mobile data, Pervasive and Mobile Computing, v.9 n.6, p.823-837, December, 2013