A Predictive Analytical Study on Factors Enhancing Customer Acquisition and Retention

Sahil Dudeja★ Rinku Dixit★★ Shailee Choudhary★★★

Abstract

CRM (Customer Relationship Management) Systems have long been used for strengthening relationships with customers thereby ensuring retention and enhancing business. Data stored in the CRM software can be analyzed to provide deep insights into the customer behavior thus influencing future products and services. Predictive Analytics are a branch of Business Analytics that helps in analyzing the current data, with the help of statistical tools, data mining algorithms, modelling tools, AI or machine learning, to make effective predictions for the future. This paper studies the impact of predictive analytics applied onto the CRM data of the sample Organization (name concealed owing to secrecy issues), which is among the front runners in the Instrumentation Industry in India and has been providing best quality Instruments and allied services through leading edge global technology. This paper examines the significant factors which help in winning a deal by using logistic regression in the reference Organization. Data are obtained from the Customer Relationship Management software provided by the company. The results presented in this paper confirm that the CRM data can be used to predict the probability of winning a deal. It also helps to find factors which are impacting 'Win' or 'Loss' of the opportunity/deal so that businesses can take precautionary measures to avoid potential loss of opportunity. Such analysis is helpful in the creation of new sales tactics, improvement of winning proportions and thereby enhancing sales.

Keywords: *CRM, Predictive Analytics, AI, Logistic Regression*

Introduction

Customer Relationship Management (CRM) is a term denoting practices, tactics and technologies used by organizations to handle customer interactions and concerned data during the customer lifecycle with the intent of enhancing customer experience and retaining them thereby increasing sales. A CRM system allows businesses to manage business relationships and data and information associated with them. With CRM systems in place, organizations can store customer information and prospective contact information, accounts, leads and sales opportunities in different locations.

CRM systems are intended to assemble data of customers that exist across channels and the various points of contacts such the company's website, call logs, chat interfaces, mailing lists, records, social media and the other marketing materials. (Mirzaei & Iyer,2014 ; Wagner et al., 2006).

The analysis of deals has traditionally relied on assigned person, location, principal code, competitor, lead source, actual CM (contribution margin) and actual revenue. Generally, assigned person is a measure of choice in winning a deal. Winning/Success rates can be potentially increased when the significant factors in winning a deal are detected through opportunity analysis and those factors are focused upon in future to win the deal/opportunity. This paper has been designed

★ Mr. Sahil Dudeja, Scholar, New Delhi Institute of Management
★★ Dr. Rinku Dixit, Associate Professor, Department of Business Analytics, New Delhi Institute of Management
★★★ Prof. Shailee Choudhary, Assistant Professor, Department of Business Analytics, New Delhi Institute of Management

to explore this aspect of the subject organization where significant factors have been identified from data stored in the CRM to predict their effect in winning or losing a deal.

Data terms used in the paper are Assigned Person referring to Location Sales Head. Principal Code refers to the partners defined by the company. Lead Sources refer to source from where the lead has been generated such as e-mail, cold calling etc. Detailed data have not been presented in this paper due to security concerns of the organization. The paper has been divide into 3 sections. In section l, authors have introduced importance of the study and choice of variables for analysis.Section 2 focusses on the literature review presenting the work done so far in this area. Section 3 focusses on the methodology used for the study. It presents data used, the models created for predicting the customer orientation. The last section presents the results and the discussions.

Literature Review

A lot of research has been undertaken in the field of CRM and the utility of the CRM data for the organization. This section provides an overview of some of the existing literatures that served as a backdrop for this study. Some of the existing literature and the focal point of their study is presented here.

Mueller in his paper written in 2010 characterized CRM as an extremely dynamic aspect of organizations' businesses. He argued that businesses would need to focus on their customers and proactively should use diverse approaches and steps for effective CRM to gain a competitive edge in their respective domains and industries.

Sinkovics and Ghauri in (2009) researched and proved the correlation between the urge to enhance engagement in CRM and variables which could be used for increase in sale such as cost of direct sales, the intensified competition in the global arena and need for information about all the of businesses in general and the behavior of the consumers.

Data mining and Knowledge Discovery in Databases (KDD) are required for picking out the models and the patterns of interests from very large databases. Fayyad and Stolorz in their paper published way back in 1997, had presented an overview of this research area, delineated the basic techniques and briefly explan that they had used some applications relating to analysis of scientific data. Out of all data mining techniques, the most popular model for customer churn has been the Logit model, which has been very effectively used for handling customer churn and thus finds use in the analysis of marketing decisions. Logistic regression is a popular model as it combines simplicity with performance and the estimated parameters are interpretable in terms of odd ratios that help in complete understanding and interpretation of the results. Further, the technique is relatively robust and popular, as is evident from its availability in almost all softwares dealing with statistical methods. In this paper the logistic regression has been applied in addition to linear regression to predict the customer orientation.

Spinello and Hames (1997) in their research find out that Wal-Mart collects the point of sale data from its various stores in various countries and sends this data to its warehouse server. Data in the warehouse server is then used for analyzing the customer buying patterns and thereby used for managing the inventory at the local store and also for identifying opportunities.

Piatetsky-Shapiro et al (1995) surveyed a the applications of data mining tools in the industries. They examined areas as fraud detection, manufacturing automation, marketing analytics etc. and how they can be deployed and adopted in the businesses. They found that these had potential values in the industry and were being researched and developed by many researchers.

Anand et al. (1996) focused on the requirements of organizations to invest in data mining solutions and tools owing to increase in data size requirements. They found that reliance on the computer programs to identify the patterns with minimal human intervention had become a necessity. They also presented a general structure of data mining which was based on the Evidence Theory that comprised methods for representing data, knowledge, data manipulation and discovery of knowledge.

Sahar F. Sabbeh (2018) in the paper compared and analyzed the efficiency of various machine learning algorithms for predicting customer churn. The techniques studied for various categories of learning, and include Discriminant Analysis, Decision Trees, instance-based learning (KNN), Support Vector Machines (SVM), Logistic Regression, and ensemble–based learning techniques.

Eva Ascarza et al. (2017) in their paper have detailed the various metrics used for measuring and monitoring customer retention. They have presented a structure for managing the retentions by exploring the emerging opportunities presented by the newest and upcoming sources of data and the latest techniques of machine learning.

Sheetal et al. (2019) in their white paper have provided an overview on how predictive analytics when applied to the big data can aid organizations in optimizing their campaigns and drives for customer acquisition and retention.

Methodology

Data used for analysis have been collected over the two years i.e 2015 to 2017 in the Customer Relationship Management software maintained by the Organization. Data contains details of all the customers and information pertaining to the won and lost opportunities.

Summary of the deals is in Table 1.

Table 1: Figures representing the Deals won and lost as per the two-year data

Total no of Deals	Number of Deals Won (Percentage of Deals Won)	Number of Deals Lost (Percentage of Deals Lost)
3572	2898 (81.13%)	694 (18.87%)

The Study of the details of the data in Table 1 (in Appendix) can help point out the significant factors that can contribute towards winning a deal and their respective impacts on predicting the success or failure of the deal. This analysis has been done in the current paper through a statistical method i.e. Logistic Regression using software R, which is known for statistical and predictive analytics.

Logistic Regression Analysis

Logistic regression is a well known and researched statistical tool for analyzing datasets that have one or more independent variables for determining the outcome. The outcome, i.e probability of winning or losing opportunities, is measured with a variable that is dichotomous in nature and has only two possible outcomes. This analysis is conducted when the dependent variable is binary and is given a binary outcome (1/0, Yes/No, True/False) when a set of independent variables exists.

Selection of Variables

The model must include all the relevant variables and it must not begin with number of variables which are more than that justified for the specified observations. More the number of variables better fit will be the model for data. But excessive variables may effect the coefficient and create an over-fit model. On the other hand a complicated model that includes several insignificant variables may have reduced predictive abilities and the results may be difficult to interpret (Yusuff 2012).

Correlation analysis is normally conducted using statistical methods to find out how two highly correlated predictors variables may lead to a

problem in Regression Analysis as they may lead to inaccuracy in the analysis. Hence, we have used VIF (Variance Inflation Factors) to remove such variables that are highly correlated to each other.

Validation

We have conducted Validation Analysis to check the suitability of logistic regression analysis. For this, data were first divided into two sets. The first set of data containing 80% samples is used as the main data and is used to finding the coefficient values. The second data set which has the remaining 20% samples is used for validating the main data. Thereafter, once the coefficient values have been obtained from the main data, the probability of each sample from the validated data are calculated. Following formula is used for calculating probability:

$$P = (Y = m) = \frac{\exp(g(x))}{1 + \exp\sum(g(x))} \qquad Eq.1$$

Probability is defined as

$$P = (Y = 0) = \frac{1}{1 + \sum\exp(g(x))} \qquad Eq.2$$

With

$$g(x) = \beta o + \sum_{p=1}^{n} \beta p X p \qquad Eq.2$$

βo is the value of the intercept coefficient and βp is the value of the coefficient for each factor that contributes to the occurrence.

Finally, the probability of each sample is cross-validated with the observed probability. With cross-validation, the percentage of correct cases obtained in the classification is thus obtained Yusuf et al. (2012).

Logistic Regression Model

Various tests, such as model fitting test, parameter estimation and classification were conducted as part of logistic regression analysis. Model fitting test helps to check whether all the variables are appropriate for usage in the logistic regression.

Following is the code in R for implementing the Model:

Model 1

```
glm(formula = target ~ ., family = binomial("logit"), data = df[, -c(1:14)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.20163  0.00003  0.00005  0.00007  2.91715

Coefficients: (4 not defined because of singularities)
```

	Estimate	Std.Error	z valve	Pr(>|z|)
(Intercept)	-4.969e+00	1.217e+00	-4.082	4.46e-05 ***
ActualCM	3.469e-08	9.270e-08	0.374	0.708279
ActualRevenue	-8.335e-08	3.387e-08	-2.461	0.013846*
LocVadodra	2.155e-01	2.339e-01	0.921	0.356943
LocDelhi	4.319e-01	2.350e-01	1.838	0.066080.
LocBangalore	-1.085e-01	2.746e-01	-0.395	0.692726
LocChandigarh	-1.561e-01	3.793e-01	-0.412	0.680633
LocChennai	1.324e+00	2.399e-01	5.518	3.42e-08 ***
LocHyderabad	6.818e-01	2.756e-01	2.473	0.013384*
LocKolkata	7.003e-01	2.922e-01	2.397	0.016546*
LocMumbai	NA	NA	NA	NA
PCodeAimilCivilAM	-5.988e-01	2.615e-01	-2.290	0.022045*
PCodeMalvernI nstrumentsLimited	1.006e+00	3.425e-01	2.936	0.003321**
PCodeOthers	-5.592e-01	2.226e-01	-2.512	0.012010*
PCodePARTICLE MEASURINGSYSTEMS	8.469e-01	3.519e-01	2.407	0.016089*
PCodeElcometer	1.091e+00	3.506e-01	3.112	0.001856**
PCodeSTRUERSAPS	-5.189e-01	4.071e-01	-1.274	0.202492
PCodeTRIMBLE	2.367e-01	2.726e-01	0.868	0.385192
PCodeAimilSMIAF	-1.632e+00	8.269e-01	-1.973	0.048498*
PCodePruftechnik	NA	NA	NA	
CompNoCompetitor	2.485e+01	3.658e+02	0.068	0.945850
CompOthers	3.922e+00	1.027e+00	3.820	0.000134***
CompEIEInstrumentsPvtLtd	4.247e+00	1.055e+00	4.025	5.71e-05***
CompHEICO	4.728e+00	1.044e+00	4.527	5.98e-06 ***
CompLeicageosystems	4.276e+00	1.110e+00	3.854	0.000116***
CompLawrenceMayo	4.204e+00	1.076e+00	3.905	9.41e-05 ***
CompIndigenousLocalmake	NA	NA	NA	NA
LSColdCall	9.647e-01	6.270e-01	1.539	0.123874
LSOther	8.676e-01	6.946e-01	1.249	0.211644
LSEmployeeReferral	4.467e-01	6.984e-01	0.640	0.522425
LSEMail	6.737e-01	6.516e-01	1.034	0.301185
LSPrincipal	9.745e-01	6.968e-01	1.398	0.161967
LSOthers	1.108e+00	6.243e-01	1.775	0.075928
LSTelephone	7.438e-01	9.684e-01	0.768	0.442439
LSTender	-7.985e-02	6.827e-01	-0.117	0.906886
LSSurvey	2.854e-01	7.333e-01	0.389	0.697091
LSCustomerContact	1.053e+00	6.733e-01	1.564	0.117932
LSAimilWebsite	NA	NA	NA	NA

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3460.0  on 3571  degrees of freedom
Residual deviance: 1634.7  on 3538  degrees of freedom
AIC: 1702.7

Number of Fisher Scoring iterations: 19
```

The model created above does not fit the equation, for model 1 stated above as part of the code, as there are a number of factors which are insignificant. Some values are showing as NA because of multicollinearity (meaning the predictors are correlated).

We need to remove some variables which are either having NA or based on z-value/p-value and remove variables with very high p-value (>.5) in each set of variables. Variance Inflation Factors (VIF) is the method for identifying collinearity amongst the explanatory variables. Higher the VIF value, the higher is the collinearity. The VIF for a single explanatory variable is obtained with the help of the r-squared value of the regression of that variable against all the other explanatory variables.

$$VIF_j = \frac{1}{1 - Rj^2} \qquad Eq.\ 4$$

In our case we remove correlated variables leaving uncorrelated set of variables for further analysis in order to avoid multicollinearity.

*VIF values >5 are commonly considered as high.

Final Model

```
glm(formula = target ~ LocVadodra + LocKolkata + ActualRevenue +
    PCodeMalvernInstrumentsLimited + PCodeElcometer + LSAimilWebsite,
    family = binomial("logit"), data = traindata)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.6362  0.3582  0.6368  0.6500  3.0333
```

Coefficients:

	Estimate	Std.Error	z valve	Pr(>|z|)
(Intercept)	1.495e+00	6.455e-02	23.158	< 2e-16 ***
LocVadodra	-7.161e-01	1.279e-01	-5.597	2.18e-08 ***
LocKolkata	7.183e-01	1.660e-01	4.327	1.51e-05 ***
ActualRevenue	-1.074e-07	2.105e-08	-5.102	3.36e-07 ***
PCodeMalvern InstrumentsLimited	9.869e-01	2.712e-01	3.639	0.000274 ***
PCodeElcometer	1.237e+00	2.660e-01	4.650	3.32e-06 ***
LSAimilWebsite	-9.226e-01	3.404e-01	-2.711	0.006717 **

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2778.8  on 2856  degrees of freedom
Residual deviance: 2639.2  on 2850  degrees of freedom
AIC: 2653.2

Number of Fisher Scoring iterations: 5
```

In the final model, all predictors have been chosen to have significant p-value (<0.01). The following observations can be made from the results.

Location-Vadodra, ActualRevenue and LeadSource-CompanyWebsite have negative impact on winning a deal whereas rest others i.e. Location-Kolkata, Principal Code-MalvernInstrumentsLimited and Elcometer have positive impact on winning a deal.

The ROC Curve presenting the results is in the figure 1:

Results

Insights from Descriptive and Statistical Analysis on data:

After successfully running the logistic regression on the above data, we ran predict function on test data to get the predicted probabilities ($0 \leq p \leq 1$). The predict function is used to calculate the predicted probability or the outcome of the categorical dependent variable which has limited number of categorical values, based on one or more independent variables. When we use the predict function on this model, it will predict the log (odds) of the Y (here, target) variable.

The default threshold score of the prediction probability is 0.5 which is the ratio of 1's and 0's in data. This decides in segregating our result into win/loss. But at times, tuning the probability cutoff can improve the accuracy in both the development and validation samples. The optimal Cut off function from Information Value package provides ways to find the optimal cut off to improve the prediction.

Based on our model, we divided the resulted probabilities into high, medium and low ranks representing the chances of winning a deal.

Below are some insights drawn from the resulted output along with the graphs to visualize the output.

a. Impact of Location on Success Rate : The impact is depicted in Figure 2

Insight- Kolkata has the highest probability of winning a deal, followed by Chennai and Mumbai. Out of all locations Chandigarh has the least chance of grabbing a deal.

b. Impact of Principal Code on Success Rate: The impact is depicted in Figure 3

Insight - Success Rate is highest when Principal Code is Civil (AM) while it is least in case of Particle Measuring Systems.

c. Impact of Lead Source on Success Rate: The impact is depicted in Figure 4

Insight - Cold call has the greatest impact on Success rate followed by Customer Contact and Email.

Conclusion

As a result, we could find out that the Location - Kolkata, Vadodara, Principal Code – Malvern Instruments Limited, Elcometer, Website and Revenue play a significant role in winning a deal.

Predictive analytics are used to find application in the context of CRM across various industries like banking, telecommunication, retail, manufacturing, insurance and healthcare.

CRM systems are therefore offering predictive analytics to keep up with the trend. Infor CRM, Salesforce, and Microsoft have all introduced predictive analytics in their latest releases and another major CRM player, SugarCRM has one in the works.

In spite of the increasing trend of applications of predictive analytics in CRM, there is a lack of inclusive literature assessment and a classification system for it. This research provided a classification framework to fulfill this gap in the literature and guide future research. The various dimensions of CRM are customer acquisition, customer attraction, customer retention, customer development and customer equity growth. Predictive analytics are mostly applied in customer retention to predict the customer churn challenges in organizations and to make knowledgeable decision to avoid these problems. In this research predictive analytics have been applied to find the significant factors to enhance customer retention and customer development.

The most predominant predictive technique, logistic regression has been applied in this research. Statistical and advanced tool 'R' is used for programming. The charts are developed in Microsoft Excel only to maintain their simplicity.

References

Anand, S. S., Bell, D. A., & Hughes, J. G. (1996). EDM: a general framework for data mining based on evidence theory. Data and Knowledge Engineering. 18, 189–223.

Ascarza, Eva , Neslin, Scott A. , Netzer, Oded , Anderson, Zachery , Fader, Peter S. , Gupta, Sunil , (2017), "In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions," Customer Needs and Solutions, available at https://doi.org/10.1007/s40547-017-0080-0.

Kamakura Wagner, Mela Carl, Ansari Asim, Bodapati Anand, Fader Pete, Iyengar Raghuram, Naik Prasad, Neslin Scott, Sun Baohong, Verhoef Peter, Wedel Michel, Wilcox Ron. (2006).Choice models and

customer relationship management. Marketing Lett.16(4):279–291

Mirzaei,Tala andIyer,Lakshmi.(2014). Application of predictive analytics in customer relationship management: A literature review and classification,Proceedings of the Southern Association for Information Systems Conference, Macon, GA, USA March 21st–22nd

Mueller, B. (2010). Dynamics of International Advertising: Theoretical and Practical Perspectives. Peter Lang second edition 2010.

Piatetsky-Shapiro, G. 1995. Knowledge Discovery in Personal Data vs. Privacy - a Minisymposium. IEEE Expert: Intelligent Systems and Their Applications. 10(2): 46-47.

Sahar F. Sabbeh (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. (IJACSA) International Journal of Advanced Computer Science and Applications, 9(2), 273-281.

Sheetal Kumari, Renu Balyan, Ashish Bhardwaj (2019). Driving Customer Acquisition and Retention with Predictive Analytics, http://bpo.rsystems.com/whitepapers/RSI-BPO-White-Paper-Driving-Customer-Acquisition-and-Retention-with-Predictive-Analytics.pdf

Sinkovics, R.R & Ghauri, P.N. (2009). New Challenges to International Marketing. Emerald Group Publishing.

Spinello, Richard A & Bernard Hames Collection (1997). Case studies in information and computer ethics. Prentice Hall, Upper Saddle River, N.J

Usama Fayyada, Paul Stolorz, (1997). Data mining and KDD: Promise and challenges. Future Generation Computer Systems. 13 (2-3): 99-115.

Yusuff H., Mohamad N., Ngah U.K. & Yahaya A.S. (2012). Breast Cancer Analysis Using Logistic Regression. International Journal of Research and Reviews in Applied Sciences. 10(1): 14-22

http://r-statistics.co/Logistic-Regression-With-R.html

https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

https://www.techadv.com/blog/3-predictive-analytics-crm

shodhganga.inflibnet.ac.in/bitstream/10603/11075/6/06_chapter2.pdf

Appendix

Figure 1: The ROC Curve presenting the results

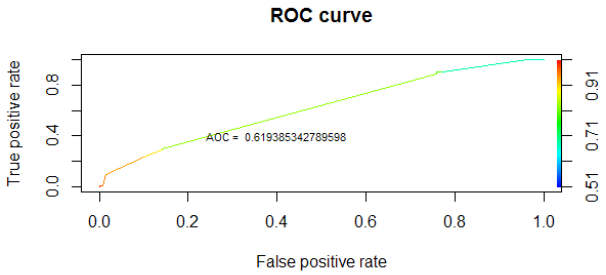

Figure 2: Impact of Location on Success Rate

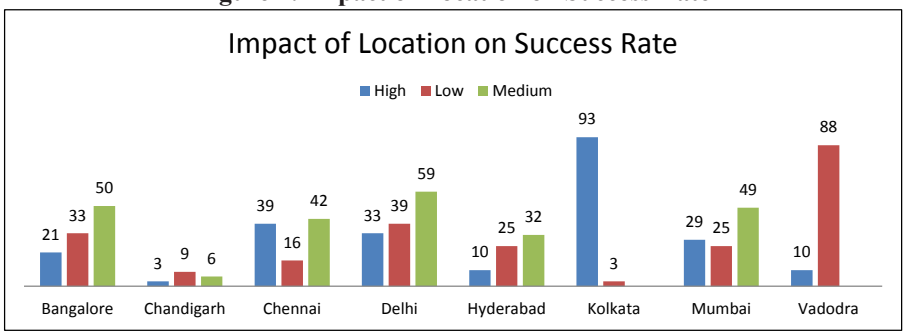

Figure 3: Impact of Principal Code on Success Rate

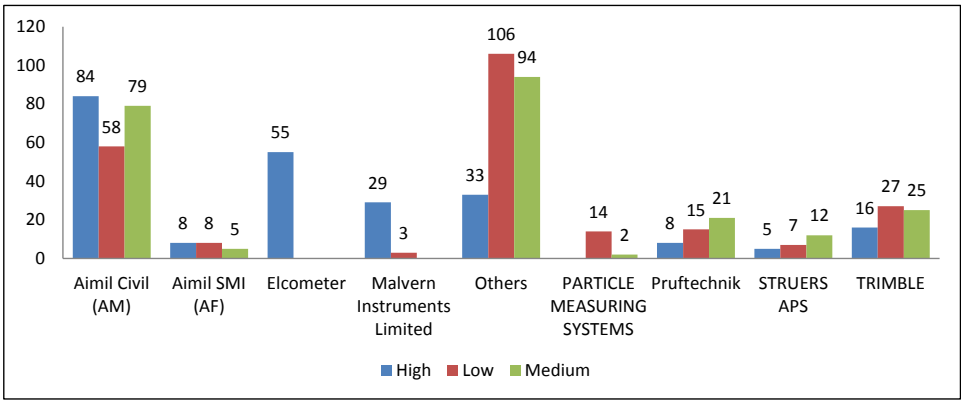

Figure 4: Impact of Lead Source on Success Rate