# RESPONSE MODELLING USING SUPPORT VECTOR MACHINES & NAÏVE BAYES

**Kanchan Wali**
Student, PGDM-Business Analytics
REVA University
Business Analyst, Analytics Edge Pvt Ltd.
Bangalore, India
Kanchanwali.BA01@reva.edu.in

**Mutturaj Baradol**
Student, PGDM-Business Analytics
REVA University
Senior Engineer, Rakuten India Inc.
Bangalore, India
Mutturajb.BA01@reva.edu.in

## ABSTRACT

Predictive analytics has become the crucial differentiating factor for highly competitive online retail companies in creating optimised solutions to provide the "right offer, right person, right time" to their customers. Response modelling is one of the important predictive modelling techniques used to get insights into the responses or behaviour of the events like repeat purchase by customers. It uses data mining techniques to find similarities between respondents from previous historical data to predict who is likely or not likely to respond in the near future.

Our experiment involves supervised learning methods, application of a Support Vector Machine (SVM) and Naïve Bayes are used to evaluate the data set for repeat purchase behaviour for an e-commerce business. We use Naïve Bayes as a baseline model and further we extend our work to more advanced classifier technique like SVM to evaluate the performance accuracy. The support vector machine is called as a linear classifier which creates a hyperplane that separates positive samples from negative samples.

The datasets used for this paper contains transaction data for an online retail shop. The customer level RFM (Recency, Frequency and Monetary) data is extracted for analysis. This paper also provides various evaluation measures for response models in terms of predictive accuracy, computational efficiency, and mid-classification error. The results of the experiments will be discussed in this paper enabling the retailer to deploy the right models with superior performance.

***Keywords:*** *Response Modelling, Support Vector Machine, Naïve Bayes, Performance accuracy.*

# 1. INTRODUCTION

Customer relationship management (CRM) is a vital differentiator enabling companies to build long-term, profitable relationships with specific customers (Ling & Yen, 2001). CRM systems and strategies are built on customer lifecycle data. The massive growth in information technologies and penetration of internet has greatly increased the opportunities for marketing to specific customer groups. Companies have transformed the way relationships between companies and their customers are managed (Ngai, 2005). One of the key components of CRM is RFM which is a method used for analysing customer value. RFM has gained popularity in the recent year and has been widely used in all spheres of marketing especially in online, database and direct marketing and has received particular attention in retail and professional services industries.

This paper presents a comprehensive review of Response modelling a technique which is one of the important predictive model technique to predict whether the customer is likely to revisit or not so that retail stores can reduce the investment on the customer retention campaigns. RFM stands for the three dimensions:

- Recency – How recently the customer has purchased?
- Frequency – How often the customer purchase?
- Monetary Value – How much money the customer spent?

# 2. LITERATURE REVIEW

Companies are increasingly deluged by data and sophisticated data mining techniques are available to marketers (Ngai et al., 2009). In order to narrow down the extant literature, the review is done on four broad topics; ‘‘Customer Relationship Management”, “Data Mining”, “Naïve Bayes”, “Support Vector Machine”. According to (Swift, 2001; Parvatiyar et al., 2001; Kracklauer et al., 2004)

The four dimensions of Customer Relationship Management are,

- Customer Identification
- Customer Attraction
- Customer Retention
- Customer Development

Overall, a CRM system helps organizations to better segregate and allocate the optimal resources to the right customer groups based on their importance. CRM tracks the customer lifecycle data including customer identification, attraction, acquisition, and retention and customer development.

CRM starts with customer identification, which is to target the prospective customers who are most likely to become customers. It also helps to identify the most profitable customers who have the higher propensity to churn because of the competition and how they can be won back (Kracklauer et al., 2004). Customer identification includes targeting the right customers, analysing them and segmentation based on future profitability. Targeting the right customer is all about finding the most profitable segments through customer analysis based on their characteristics. Customer segmentation strategies would subdivide the entire customer group into smaller customer subgroups or segments based on their similarity in characteristics. (Woo et al., 2005).

The next crucial component of the CRM process is customer attraction. Based on various promotional activities the companies identify and woo these customers and convince them to visit their website or store or channel.  Multiple promotions motivate customers to place orders through various channels. Customer retention is another key area of focus through which the companies ensure that right customers are loyal and do not churn due to competitive activities. The life cycle management also includes customer development, which in turn creates more upselling and cross-selling opportunities for customers (Cheung et al., 2003; Liao & Chen, 2004; Prinzie & Poel, 2005).

The Customer management system is considered as a closed cycle of these dimensions (Au & Chan, 2003; Kracklauer et al., 2004; Ling & Yen, 2001). All four dimensions have a common goal of creating an in-depth understanding of customers so that customer value can be maximized. Such a goal can be met with the Data mining techniques by mining unseen customer characteristics and behaviours from large databases. The data mining consists of building a model using the data (Carrier & Povel, 2003). Generally, the following types of models are built on the data.

1. Association
2. Classification
3. Clustering
4. Forecasting
5. Regression
6. Sequence discovery
7. Visualization.

Various authors have highlighted these models in their articles (Ahmed, 2004; Carrier & Povel, 2003; Mitra, Pal, & Mitra, 2002; Shaw et al., 2001; Turban et al., 2007). Each modelling techniques use various machine learning techniques. Typically the choice of data mining techniques should be based on the data characteristics and business requirements (Carrier & Povel, 2003). Ngai et al.,  (2009) depicted a few of the widely used data mining algorithms like  Association rules, Decision trees, Neural networks, *k*-Nearest neighbour, Naïve Bayes, Linear/logistic regression modelling and Support Vector Machines for each phase in the CRM (Diagram 1).

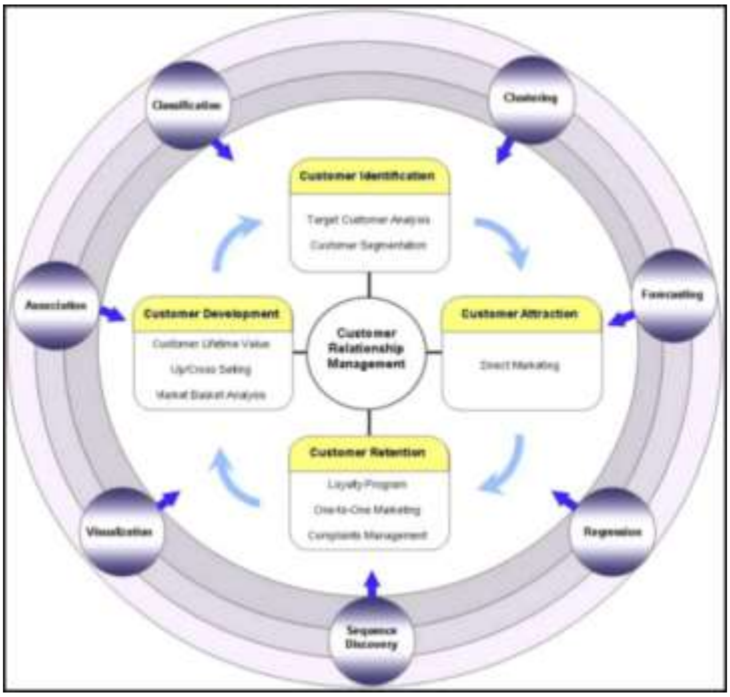
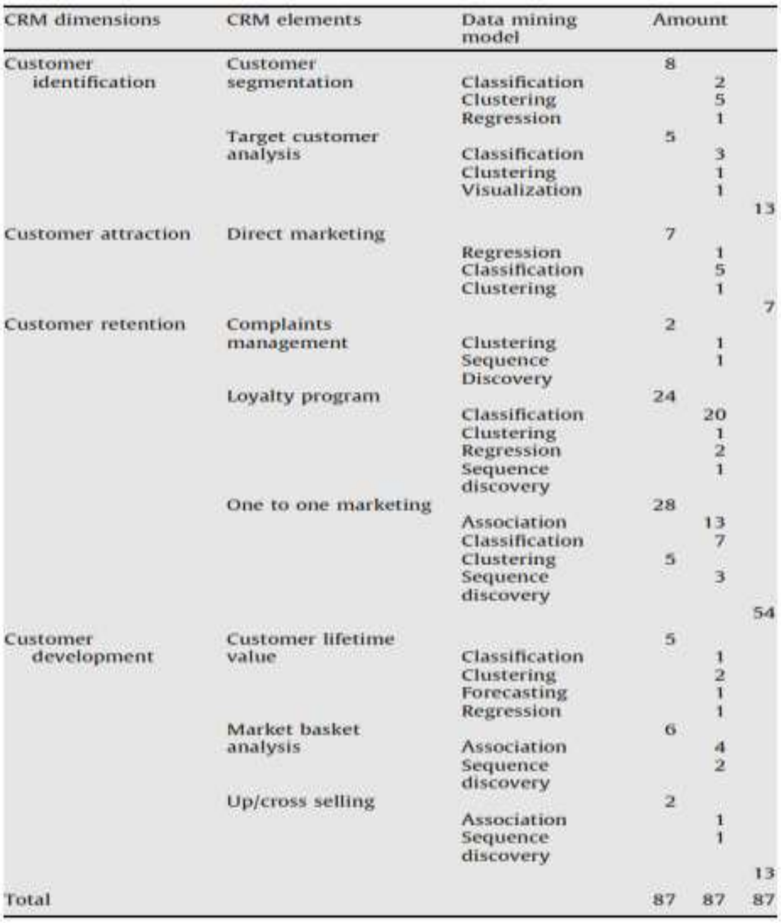


**Diagram 1** CRM Model (Ngai et.al., 2009)

In recent years, there is a surge of articles written on applying various Machine Learning Techniques for various stages of CRM (Table 1). The classification models are extensively used in target customer analysis as per the Ngai et al., (2009).  Shin et al., (2009) suggested applying Support Vector Machine (SVM) in direct marketing but, also lists the practical difficulties of applying SVM like large training data, class imbalance and scoring from binary SVM output. Ayetiran et al., (2012) uses a predictive response model with the Bayesian algorithm. In the article, the authors have developed a Naïve Bayes algorithm data mining techniques to predict the probability of a customer in Ebedi Microfinance bank will respond to a promotion or an offer.

Companies use the knowledge of consumer behaviour to segment and design marketing strategies, and measure marketing performance (Schiffman & Kanuk, 1991). Techniques like SVM is rarely applied in CRM and to build customer response model, with exceptions (Viaene et al., 2001). The main purpose of response modelling is to improve future return on investment on marketing (Shin & Cho, 2006). Response models have been proven to be a highly profitable tool in fine-tuning marketing strategies (Elsner et al., 2004). Though hardly applied in real life situations, SVMs have great generalisation ability and have strong performance scores compared to traditional modelling approaches (Cui & Curry, 2005). Coussemet & Poel (2006) have used SVM in a newspaper subscription contest, and have proved that SVM has good generalisation ability when compared to logistic regression and random forest. (Prasad, & Anjaneyulu (2015) makes a comparative study of SVM and logistic regression and suggests SVM is a better classifier.



| CRM dimensions | CRM elements | Data mining model | Amount | | |
|---|---|---|---|---|---|
| Customer identification | Customer segmentation | | 8 | | |
| | | Classification | | 2 | |
| | | Clustering | | 5 | |
| | | Regression | | 1 | |
| | Target customer analysis | | 5 | | |
| | | Classification | | 3 | |
| | | Clustering | | 1 | |
| | | Visualization | | 1 | |
| | | | | | 13 |
| Customer attraction | Direct marketing | | 7 | | |
| | | Regression | | 1 | |
| | | Classification | | 5 | |
| | | Clustering | | 1 | |
| | | | | | 7 |
| Customer retention | Complaints management | | 2 | | |
| | | Clustering | | 1 | |
| | | Sequence Discovery | | 1 | |
| | Loyalty program | | 24 | | |
| | | Classification | | 20 | |
| | | Clustering | | 1 | |
| | | Regression | | 2 | |
| | | Sequence discovery | | 1 | |
| | One to one marketing | | 28 | | |
| | | Association | | 13 | |
| | | Classification | | 7 | |
| | | Clustering | 5 | | |
| | | Sequence discovery | | 3 | |
| | | | | | 54 |
| Customer development | Customer lifetime value | | 5 | | |
| | | Classification | | 1 | |
| | | Clustering | | 2 | |
| | | Forecasting | | 1 | |
| | | Regression | | 1 | |
| | Market basket analysis | | 6 | | |
| | | Association | | 4 | |
| | | Sequence discovery | | 2 | |
| | Up/cross selling | | 2 | | |
| | | Association | | 1 | |
| | | Sequence discovery | | 1 | |
| | | | | | 13 |
| Total | | | 87 | 87 | 87 |

**Table 1** CRM dimensions and data mining model and no. of papers (Ngai et al., 2009)

## 3. METHODOLOGY

The data used for the research is from Jim Porzak’s repository (https://github.com/ds4ci/rfmr), where the customer orders are from an online and catalogue retailer using R and R Studio. RFM (recency, frequency, monetary) analysis is a marketing technique used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchased (frequency), and how much the customer spends (monetary). RFM analysis is based on the marketing axiom that "80% of your business comes from 20% of your customers” (Techtarget.com).

The customer transaction dataset has 135k unique customers corresponding to these customers there were 541k orders in the dataset for ~900 products in 5 categories over 2 ½ years. The description of the dataset is shown in Table 2.

| | |
|---|---|
| **OrderID** | The unique order identifier |
| **OrderDate** | The date of the order |
| **OrderChannel** | How the order came in (phone1, phone2, web1, web2) |
| **SKU_ID** | Stock Keeping Unit. The unique product identifier |
| **CustID** | The unique Customer identifier |
| **Quantity** | How many items of the SKU were ordered |
| **Amount** | The total value amount for the order line. (unit price) * Quantity |
| **Category** | The product category code (C, G, I, N, T, X) |

**Table 2** Variables in the Dataset

The original data needed to be cleaned and the customer analytical record is prepared using the DataMart which is developed using an aggregated variable like Recency, Frequency, Monetary. RFM variables are to be created with respect to each Customer ID which will help in identifying the customers who are likely to respond in the near future, normally the most recent, frequent, and high spending customers are more likely to respond than others. The distribution for the predictors R, F, M and the Visit response variable were evaluated by examining frequency tables for categorical variables and calculating the mean, standard deviation, minimum and maximum values for quantitative variables as shown in Table 3.

| CustID | | Recency | | Frequency | | Monetary | | Visit | |
|---|---|---|---|---|---|---|---|---|---|
| Minimum | 7 | Minimum | 0 | Minimum | 3 | Minimum | 1 | 0 | 43921 |
| Median | 205841 | Median | 18 | Median | 60 | Median | 2 | 1 | 1696 |
| Mean | 281905 | Mean | 18.72 | Mean | 80.51 | Mean | 2.55 | | |
| Maximum | 503309 | Maximum | 32 | Maximum | 7531.6 | Maximum | 98 | | |

**Table 3** Descriptive Statistics of the DataMart

The RFM variables are calculated using CustID, OrderID, Amount variables for the dataset. The Recency is calculated using the CustID who has visited lately to the store, Frequency is the length of CustID and OrderID, Monetary is the sum of all Amount by CustID.

The response model data mart is prepared on CustID as an identifier and RFM variables, the data is considered for last seven months and the responses for six months is calculated on the last to last month. The responses will be as a categorical variable in terms of visit the store or not (0 or 1). The Visit variable is 1 only if that particular customer has visited the store in past six months irrespective of his purchase behaviour else it will be 0.

| CustID | Recency_6 | Frequency_6 | Monetary_6 |
|---|---|---|---|
| 7 | 7 | 2 | 19 |
| 13 | 13 | 3 | 25 |
| 16 | 16 | 1 | 25 |
| 24 | 24 | 2 | 30 |
| 25 | 25 | 5 | 31 |
| 29 | 29 | 1 | 30 |

**Table 4** A snapshot of RFM based on customer ID for 6 months

The sampling of the data is carried out using the Hold-Out method and the train and test data is used for modelling. Classification models like Naïve Bayes and SVM are then applied to see the responses. The validation of the model is done through Model accuracy, precision rate and misclassification rate.

| CustID | Recency | Frequency | Monetary | Visit |
|---|---|---|---|---|
| 7 | 7 | 2 | 19 | 0 |
| 13 | 13 | 3 | 25 | 0 |
| 16 | 16 | 1 | 25 | 0 |
| 24 | 24 | 2 | 30 | 0 |
| 25 | 25 | 5 | 31 | 0 |
| 29 | 29 | 1 | 30 | 0 |

**Table 5** Categorical Variable -visited the store or not in terms of (0 or 1)

The other validation techniques for classification cannot be used because the responses were imbalanced. The R scripts and final models are uploaded to GitHub. The link to access the script is here.

## 4. FINDINGS

Using Classification techniques like Naïve Bayes and Support Vector Machine for the modelling. The two models are trained on the Train data and validated for the Test data. The comparative analysis of the study shows these results (Table 6, 7 and 8).

| **Naive Bayes** | | |
|---|---|---|
| Predicted | Actual | |
| | 0 | 1 |
| 0 | 17409 | 533 |
| 1 | 159 | 145 |
| Accuracy | | 0.96 |
| Precision | | 0.21 |
| Error | | 4% |

| **Support Vector Machine** | | |
|---|---|---|
| Predicted | Actual | |
| | 0 | 1 |
| 0 | 17349 | 235 |
| 1 | 219 | 443 |
| Accuracy | | 0.97 |
| Precision | | 0.65 |
| Error | | 3% |

**Table 6** SVM and NB Comparative Analysis

The Naïve Bayes and Support Vector Machines are a very good classifier. Both the classifiers are classifying true visitors and non-visitors equally.

| | |
|---|---|
| Visiting | 678 |

**Table 7** Actual visitor details in test data

Out of 18246 actual visitors in the test data, the total respondents are 678, Naïve Bayes classifies true visitors as 304 and SVM as 662.

| Naïve Bayes | |
|---|---|
| Visiting | 304 |

| SVM | |
|---|---|
| Visiting | 662 |

**Table 8** Classifying true visitors

With this comparative analysis, SVM is a better classifier for responders. The SVM suggests that 443 customers are loyal to the stores and 235 customers are prone to attrition, so efforts like investment could be done in DMA to retain them to the stores.

## 5. CONCLUSION

The paper attempts to find the right analytical technique to find the customers who are more frequent buyers and targeting only those valuable customers who are important for any retailer to reduce investment in the customer retention campaigns. Compared to Naïve Bayes classifier, SVM seems to be a good classifier based on the above results. The SVM classifier suggests that there are loyal customers visiting the store and additionally we can increase the footfall of the customers knowing their purchase behaviour. The additional customers can be identified using SVM. The investment of time and money could be made only on these customers and saving the overall loses. The model also identifies the customers who are actually prone to attrition, which would reduce the investment and costs on retailers. Hence looking at the techniques and results this model could be used in any of the domains like retail, banking, telecom, hospitality etc.

## REFERENCES

Au, W. H., & Chan, K. C. C. (2003). Mining fuzzy association rules in a bank-account database. IEEE Transactions on Fuzzy Systems, 11, 238–248.

Ayetiran, E. F., & Adeyemo, A. B. (2012). A data mining-based response model for target selection in direct marketing. International Journal of Information Technology and Computer Science (IJITCS), 4(1), 9.

Ahmed, S. R. (2004, April). Applications of data mining in the retail business. In Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on (Vol. 2, pp. 455-459). IEEE.

Carrier, C. G., & Povel, O. (2003). Characterising data mining software. Intelligent Data Analysis, 7, 181–192.

Cheung, K. W., Kwok, J. T., Law, M. H., & Tsui, K. C. (2003). Mining customer product ratings for personalized marketing. Decision Support Systems, 35, 231–243.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert systems with applications, 34(1), 313-327.

Elsner, R., Krafft, M., & Huchzermeier, A. (2004). Optimizing Rhenania's direct marketing business through dynamic multilevel modelling (DMLM) in a multi-catalogue-brand environment. Marketing Science, 192-206.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Customer management as the origin of collaborative customer relationship management. Collaborative Customer Relationship Management - taking CRM to the next level, 3–6.

K Prasad, G Anjaneyulu, (2015) A Comparative Analysis of Support Vector Machines & Logistic Regression for Propensity Based Response Modelling- International Journal of Business.

Ling, R., & Yen, D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. Journal of Computer Information Systems, 41, 82–97.

Liao, S. H., & Chen, Y. J. (2004). Mining customer knowledge for electronic catalogue Marketing. Expert Systems with Applications, 27, 521–532.

Margaret Rouse, Techtarget.com
http://searchdatamanagement.techtarget.com/definition/RFM-analysis (Last accessed on 23/10/2017)

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. IEEE Transactions on Neural Networks, 13, 3–14.

Ngai, E. W. T. (2005). Customer relationship management research (1992–2002): An academic literature review and classification. Marketing Intelligence, Planning, 23, 582–605.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

Parvatiyar, A., & Sheth, J. N. (2001). Customer relationship management: Emerging practice, process, and discipline. Journal of Economic & Social Research, 3, 1–34.

Prasad, K. V. N. K., & Anjaneyulu, G. V. S. R. (2015). A Comparative Analysis of Support Vector Machines & Logistic Regression for Propensity Based Response Modeling. International Journal of Business Analytics and Intelligence, 3(1), 7.

Prinzie, A., & Poel, D. V. D. (2005). Constrained optimization of data-mining problems to improve model performance: A direct-marketing application. Expert Systems with Applications, 29, 630–640.

R Kohavi, R Parekh, Visualizing RFM segmentation (2004) - Proceedings of the 2004 SIAM International, 2004 – SIAM

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. Decision Support Systems, 31, 127–137.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 28(1), 127-135.

Shin, H., & Cho, S. (2007). Neighborhood property–based pattern selection for support vector machines. Neural Computation, 19(3), 816-855.

Schiffman, L. G. Kanuk. LL, (1991). Consumer Behavior, Englewood Cliffs, PrenticeHall Inc.

Swift, R. S. (2001). Accelerating customer relationships: Using CRM and relationship technologies. Prentice Hall Professional.

Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. (2007). Decision support and business intelligence systems (Eighth ed.). Pearson Education.

Viaene, S., & Cumps, B. (2005). CRM Excellence at KLM Royal Dutch Airlines. Communications of the Association for Information Systems, 16(1), 27.

Woo, J. Y., Bae, S. M., & Park, S. C. (2005). Visualization method for customer targeting using customer map. Expert Systems with Applications, 28763–772.