# Analysis of customer relationship management using Machine Learning

**Dr. V.Anandi[a], Dr. M. Ramesh[b]**

[a] Associate Professor of ECE, M S Ramaiah Institute of Technology (Autonomous), Bangalore
[b]Professor of Management Studies, Alliance University,Bangalore

**Abstract:**Customer retention is key to business sustenance and have become more important in the quality of service (QoS) that organizations can provide them. Services provided by different vendors are not highly distinguished which increases competition between organizations to maintain and increase their QoS. Customer Relationship Management systems are used to enable organizations to acquire new customers, establish a continuous relationship with them and increase customer retention for more profitability Understanding of customer satisfaction, persona analysis of customer website visit pattern, customer feedback analysis are requirement for future to retain customers. This paper discusses methods using machine learning and analysis to achieve these objectives, and redesign the product according to customized needs.CRM systems use machine-learning models to analyze customers' personal and behavioural data to give organization a competitive advantage by increasing customer retention rate. Those models can predict customers who are expected to churn and reasons of churn. Predictions are used to design targeted marketing plans and service offers. This paper tries to compare and analyze the performance of different machine-learning techniques that are used for prediction problem. Analytical techniques that belong to different categories of learning are chosen for this study. The chosen techniques include Discriminant Analysis, Decision Trees (CART), instance-based learning (k-nearest neighbours), Support Vector Machines, Logistic Regression, ensemble–based learning techniques (Random Forest, Ada Boosting trees and Stochastic Gradient Boosting), Naïve Bayesian, and Multi-layer perceptron. Models were applied on a dataset of companies that contains CRM feedback records. Results show that both random forest and ADA boost outperform all other techniques with almost the same accuracy 97%. Both Multi-layer perceptron and Support vector machine can be recommended as well with 95% accuracy. Decision tree achieved 92%, naïve Bayesian 90% and finally logistic regression and Linear Discriminant Analysis (LDA) with accuracy 88.7%.
**Keywords:** Customer relationship management (CRM), customer retention, Analytical CR,; Business Intelligence, Machine-Learning, Predictive Analytics, Data Mining, Customer Churn.

## 1. Introduction

This research is about the organizational after-sales application, through web application a part of the online platform, supported by more than 13 International languages and more than 190+ countries. Currently, the website examined is integrated with the products and services of group companies. This website has per day almost more than 3 million visitors including authenticated and anonymous experiences. Users can search their assets making use of this website. The product's details, warranty, the system configuration, are visible to the users through the service tag of this website. Utilizing this service tag, users do the online diagnostic if it's in the warranty provided by the user organization with auto dispatch too. Users can download and install the product drivers and can upgrade to the operating system too. Users can contact the service advisor for help using chat, email, phone for this users need to key the service tag, and base on the queue available this website will share the information and avail the facilities to chat with Agents. Using Order Number user can track order details and do the after-sales operations. This website design is a revenue saver, when a user is calling to host a support agent, chatting with an agent, or contacting using emails it requires support from hourly basic, and almost roughly it's 15 dollars per hour**.** This website is facilitating to do self-resolution of all the doubts and clarifications. For any organization customer satisfaction is important for the organization. Many factors are contributing to customer satisfaction as we are always trying easy, fast, and simplification. For achieving these goals if the website is serving data without exception it's contributing to the CSAT.

As part of the study, from raw data downloaded from IIS logs, the three quarter's web traffic of the host were analyzed. Using Splunk tools fetched the quarterly data and established the relationship with the data. Based on the analysis the solution which the company can implement to improve customer satisfaction was inferred. As we know customer satisfaction is dependents on too many factors, errors and exceptions are one of them. For addressing these pain points a strategy needs to be created . Need to categorize the erred traffic and need to address with fixes, it can be broken link, Wrong web crawler, Wrong Service request, Bot attack, code issues, Service availability, Data Setup, Response Time, Database response time and better Deployable Environment (Availability, Performance, Code Quality, Dependence Environment availability, etc) will resultant to reduce Error and exceptions will improve Customer Satisfaction, it will save revenue to the company.

## 2.Review Of Related Studies

**Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2009)** discuss about Customer relationship management (CRM) to build relations with the most profitable clients by performing customer segmentation and designing appropriate marketing tools. Several statistical techniques have been applied for market segmentations. In this article, a three-stage methodology is proposed that combines marketing feature selection, customer segmentation through univariate and oblique decision trees, and a new CPA function based on marketing, data warehousing, and opportunity costs linked to the analysis of different scenarios.**QAS. (2006, September).**Business data decays at a rapid rate, arguably, faster even than consumer data. Also discusses how much B2B direct mail reaches its intended recipient, and how much of it is deemed relevant.The survey highlights the challenge facing database professionals to keep their B2B customer and prospect data up to date. **Kim, Y. , Street, W.N. , Russell, G.J. , &Menczer, F. (2005)** conducted Principal component analysis (PCA) of customer background information followed by logistic regression analysis of response behaviorfor database marketers. In this paper, we propose a new approach that uses articial neural networks (ANN's) guided by genetic algorithms (GA's) to target households. We show that the resulting selection rule is more accurate and more parsimonious than the PCA/logit rule when the manager has a clear decision criterion**He, Z. , Xu, X. , Huang, J.Z. , & Deng, S. (2004)** conducted a study on Outliers, or commonly referred to as exceptional cases, exist in many real-world databases.In this paper, they consider the *class outlier detection problem* ‘given a set of observations with class labels, find those that arouse suspicions, taking into account the class labels’.They have developed the notion of class outlier and propose practical solutions by extending existing outlier detection algorithms to this case. Furthermore, its potential applications in CRM (customer relationship management) are also discussed. Finally, the experiments in real datasets show that their method can find interesting outliers and is of practical use.

## 3.Objectives Of The Study

While there are many factors to improve customer satisfaction, the immediate objectives to focus on this study are mentioned below.

- Understanding of customer satisfaction as persona.
- Analysis of customer website visit pattern and Future to make customer.
- 360° customer feedback analysis.
- Requirement for future to retain customer using machine learning.

## 4.Hypotheses Of The Study

- Awareness on Customer Relationship Management to build relations with most profitable clients.
- To gain a better understanding of the impact of business data decay and its cost.
- Principal component analysis (PCA) of customer background information followed by logistic regression analysis
- Find interesting outliers of practical use.

## 5.Statistical Techniques Used in the Present Study

**R language for Multiple Linear Regression**

## 6.Methodology: Details of the data Collected

The data for this analysis was procured through Internet access of host website’s Microsoft Internet Information Services (IIS) traffic logs, These Microsoft Internet Information Services (IIS) traffic logs are website data collected from 80 User interface server and 16 Service Servers. For analyzing these data, used analytic R language and Excel. Using R language, descriptive and predictive analysis were used, including data display, multiple linear regression and hypothesis analysis. Excel package was used to do data display and analysis. The data contains web site traffic information for a period of three quarters during January to December 2019. The following columns were available:

- Host name
- Microsoft Internet Information Services (IIS) status code
- Quarter periods

- Time taken
- Used URL
- Browser version
- Language used

### 6.1 Data Preparation

At the start, the data preparation includes analysis with primary data, through which was able to organize and view basic data on Microsoft internet information services (IIS) website traffic and quality of traffic means healthy traffic (IIS response code 2xx - succes) and errored traffic (IIS response code 4xx - client error and 5xx - server error) and did the analysis of three-quarter data, using R language for correlogram, multiple regression and Excel for bar chart.

### 6.2 Data Analysis and Interpretation

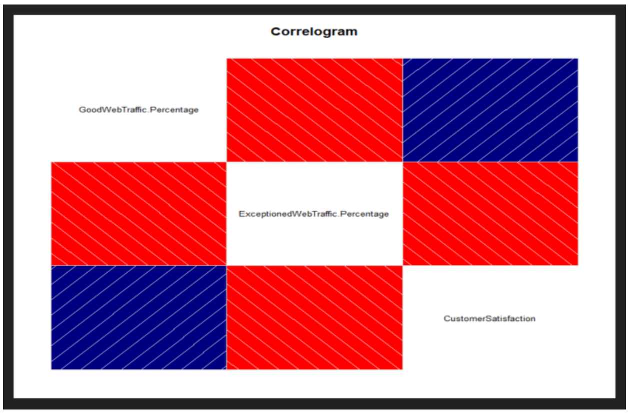
The analysis and the findings discussed here towards the first business question that the organization had: - Identifying and arriving at the right catalyst for Customer Satisfaction (CSAT) and the revenue saved.The business question was prolonged: Define the matrix for CSAT is increasing the good traffic is positively affecting CSAT means addressing the website error and exceptions. The analysis of the web site traffic was carried out as excel which offered the following data cuts based on which the graphs were produced. Based on the above, we will able to identify web site traffic that contributes most of the good traffic and those saved the revenue and increased the CSAT. Post this, combining both of these data could provide us the view of the CSAT contributor.For this, we ran series of R language query using multiple liner regression and used some excel analysis.

### 6.3Findings and Discussions

**Used R language to determine the relation of data using correlogram and** the output is shown in Figure 1.

```
install.packages("ggplot2")
library(ggplot2)
Project Data=read.csv(file.choose())
# Correlogram
install. packages("ggplot2")
library(ggplot2)
install. packages("corrgram")
library(corrgram)
corrgram (ProjectData, order=NULL, panel=panel.shade, text.panel=panel.txt,    main="Correlogram")
```

Figure 1 : Output using correlogram (Source: Generated using R Program Console.)




This means Customer Satisfaction is highly positive correlated with Good Web traffic and Customer Satisfaction is highly negative correlated with Error Web traffic.

**Used R language for Multiple Linear Regression**

**# Multiple Regression**

```
install.packages("usdm")
library("sp")
library("raster")
```

library("usdm")
head(ProjectData)
**# Dependent variable will be Customer Satisfaction**
Data Frame_Project Data=data.frame (Project Data $ Good Web Traffic. Percentage, Project Data$ Exception Web Traffic. Percentage, Project Data$ Customer Satisfaction)
Cor (DataFrame_Project Data)
ResultLM_ProjectData=lm(ProjectData$CustomerSatisfaction~ProjectData$GoodWebTraffic.Percentage+ProjectData$ExceptionedWebTraffic.Percentage)
ResultLM_ProjectData
Summary (ResultLM_ProjectData)
Below is the Output of Multiple Linear regression Summary.

```
# Call:
#   lm(formula = ProjectData$CustomerSatisfaction ~ ProjectData$GoodWebTraffic.Percentage +
#        ProjectData$ExceptionedWebTraffic.Percentage)
#
# Residuals:
#   1       2       3
# -0.1341  0.3900 -0.2559
#
# Coefficients: (1 not defined because of singularities)
# Estimate Std. Error t value Pr(>|t|)
# (Intercept)                                  -1176.17     122.26   -9.62    0.0659 .
# ProjectData$GoodWebTraffic.Percentage          12.521      1.241   10.101    0.0628 .
# ProjectData$ExceptionedWebTraffic.Percentage     .001       .001     .001    0.0001
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.4853 on 1 degrees of freedom
# Multiple R-squared:  0.9903,  Adjusted R-squared:  0.9806
# F-statistic:   102 on 1 and 1 DF,  p-value: 0.06283
```

As per above output adjusted R-square has 98%, and P value is near to 0.05.
Blow is the hypothesis for this:

```
#Hypothysis for liner regression
#Ho : b1=0(GoodWebTraffic and  ExceptionedWebTraffic does not influance  CustomerSatisfaction ) (Null Hypothesis)
#Ha : b1!=0(GoodWebTraffic and  ExceptionedWebTraffic influancing   CustomerSatisfaction )(Alternate Hypothesis)
```

Hence Null hypothesis is satisfying.

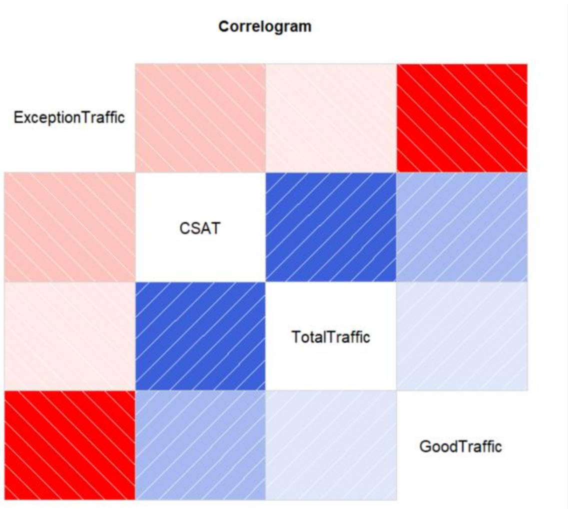
## Analysis for weekly data:

**# Correlogram ,below is the R code analysis and diagram.**

- Exceptions and Good Traffic have highly negative relation.
- CSAT and positive relation with good traffic.

install.packages("ggplot2")
library(ggplot2)
install.packages("corrgram")
library(corrgram)
corrgram(ProjectData, order=NULL, panel=panel.shade, text.panel=panel.txt,
main="Correlogram")

**Figure 2:** # Correlogram analysis weekly data using R Code



**# Used R language for Multiple Linear Regression**
**Dependent variable will be Customer Satisfaction with Exception traffic with total traffic**
DataFrame_ProjectData=data.frame(ProjectData$GoodTraffic,ProjectData$ExceptionTraffic,ProjectData$CSAT)
cor(DataFrame_ProjectData)
vif(DataFrame_ProjectData[,1:2])
**ResultLM_ProjectData=lm(ProjectData$CSAT~ProjectData$ExceptionTraffic+ProjectData$ï..TotalTraffic)**
ResultLM_ProjectData
summary(ResultLM_ProjectData)
Below is the Output of Multiple Linear regression Summary.

```
# Call:
#  lm(formula = ProjectData$CSAT ~ ProjectData$ExceptionTraffic +
#      ProjectData$ï..TotalTraffic)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -5.0914 -2.2487 -0.6159  2.0898  7.5209
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)                   5.638e+01  1.294e+00  43.559  < 2e-16 ***
#   ProjectData$ExceptionTraffic -3.972e-01  5.763e-01  -0.689 0.494914
# ProjectData$ï..TotalTraffic   2.552e-08  6.559e-09   3.891 0.000402 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 2.894 on 37 degrees of freedom
# Multiple R-squared:  0.3055,  Adjusted R-squared:  0.268
# F-statistic: 8.139 on 2 and 37 DF,  p-value: 0.001177
```

Exceptional traffic have lower correlation with total traffic hence good traffic is affecting CSAT

**Table 1 : Excel to visualize the data three quarter and Financial Week data for this**.

| Week | Period | Good Traffic | Exception Traffic | CSAT |
|---|---|---|---|---|
| FY19-26 | Jul 28 - Aug 03 | 98.10 | 1.90 | 54.85 |
| FY19-27 | Aug 04 - Aug 10 | 98.32 | 1.68 | 54.09 |
| FY19-28 | Aug 11 - Aug 17 | 98.29 | 1.71 | 55.04 |
| FY19-29 | Aug 18 - Aug 24 | 98.93 | 1.07 | 51.2 |
| FY19-30 | Aug 25 - Aug 31 | 98.41 | 1.59 | 54.18 |
| FY19-31 | Sep 01 - Sep 07 | 98.94 | 1.06 | 56.19 |
| FY19-32 | Sep 08 - Sep 14 | 98.66 | 1.34 | 56.63 |
| FY19-33 | Sep 15 - Sep 21 | 99.08 | 0.92 | 52.84 |
| FY19-34 | Sep 22 - Sep 28 | 98.78 | 1.22 | 54.1 |
| FY19-35 | Sep 29 - Oct 05 | 97.79 | 2.21 | 53.87 |
| FY19-36 | Oct 06 - Oct 12 | 98.68 | 1.32 | 55.55 |
| FY19-37 | Oct 13 - Oct 19 | 96.94 | 3.06 | 54.38 |
| FY19-38 | Oct 20 - Oct 26 | 95.30 | 4.70 | 55.99 |
| FY19-39 | Oct 27 - Nov 02 | 96.05 | 3.95 | 56.84 |
| FY19-40 | Nov 03 - Nov 09 | 97.21 | 2.79 | 55.53 |
| FY19-41 | Nov 10 - Nov 16 | 98.66 | 1.34 | 55.94 |
| FY19-42 | Nov 17 - Nov 23 | 98.89 | 1.11 | 56.66 |
| FY19-43 | Nov 24 - Nov 30 | 98.61 | 1.39 | 57.82 |
| FY19-44 | Dec 01 - Dec 07 | 99.11 | 0.89 | 58.83 |
| FY19-45 | Dec 08 - Dec 14 | 99.14 | 0.86 | 59.01 |
| FY19-46 | Dec 15 - Dec 21 | 99.01 | 0.99 | 59.62 |
| FY19-47 | Dec 22 - Dec 28 | 98.64 | 1.36 | 61.88 |

| | | | | |
|---|---|---|---|---|
| FY19-48 | Dec 29 - Jan 04 | 98.08 | 1.92 | 59.67 |
| FY19-49 | Jan 05 - Jan 11 | 98.46 | 1.54 | 61.68 |
| FY19-50 | Jan 12 - Jan 18 | 97.77 | 2.23 | 61.52 |
| FY19-51 | Jan 19 - Jan 25 | 98.50 | 1.50 | 64.19 |
| FY19-52 | Jan 26 - Feb 01 | 98.08 | 1.92 | 64.57 |
| FY20-01 | Feb 02 - Feb 08 | 98.18 | 1.82 | 62.02 |
| FY20-02 | Feb 09 - Feb 15 | 99.10 | 0.90 | 61.86 |
| FY20-03 | Feb 16 - Feb 22 | 98.93 | 1.07 | 58.77 |
| FY20-04 | Feb 23 - Mar 01 | 98.58 | 1.42 | 60.08 |
| FY20-05 | Mar 02 - Mar 08 | 99.10 | 0.90 | 62.12 |
| FY20-06 | Mar 09 - Mar 15 | 98.30 | 1.70 | 62 |
| FY20-07 | Mar 16 - Mar 22 | 99.17 | 0.83 | 60.13 |
| FY20-08 | Mar 23 - Mar 29 | 98.78 | 1.22 | 61.31 |
| FY20-09 | Mar 30 - Apr 05 | 98.10 | 1.90 | 59.78 |
| FY20-10 | Apr 06 - Apr 12 | 98.84 | 1.16 | 59.32 |
| FY20-11 | Apr 13 - Apr 19 | 98.43 | 1.57 | 61.01 |
| FY20-12 | Apr 20 - Apr 26 | 98.57 | 1.43 | 60.17 |
| FY20-13 | Apr 27 - May 03 | 97.97 | 2.03 | 62.8 |

**Figure 3:** Exception traffic vs CSAT.

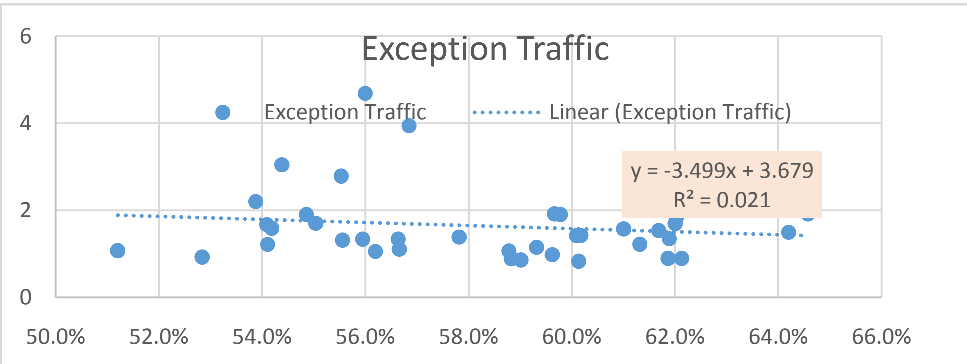



Figure 3 is the Exception traffic vs CSAT. As per Graph we can conclude CSAT is declining as there are more exceptions. The Linear equation will be y=-3.4995x+3.6791. And R square value will be .2% which is slightly correlated.

**Figure 4:**Good traffic vs CSAT

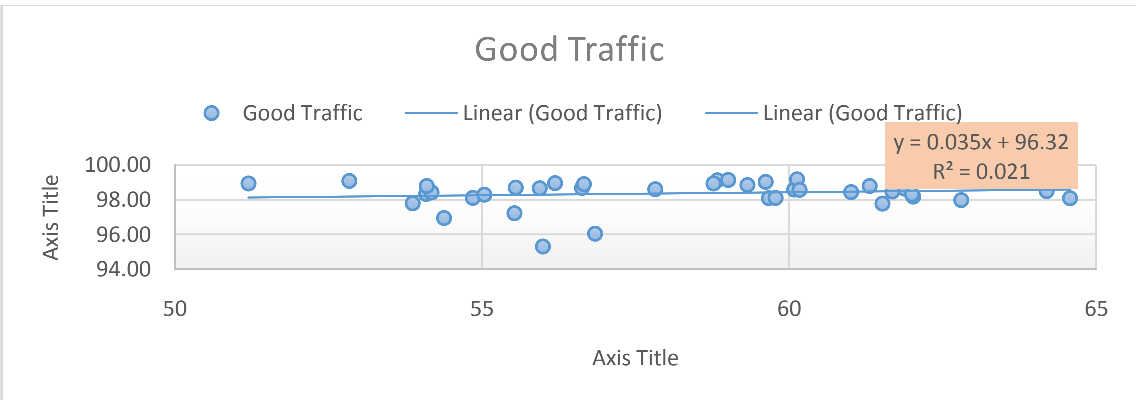


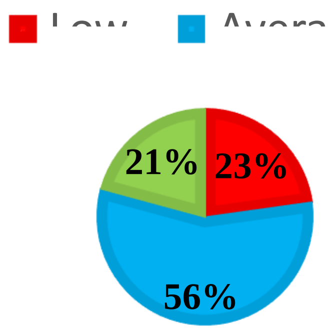
Figure 4 illustrates Good traffic vs CSAT. As per Graphical linear tend line we can conclude CSAT is improving as have increasing good traffic. Linea equation will be y=0.035x+96.322, and R square value will be .2% which means slightly correlated.

Table 2 :**Excel to visualize the data after querying from Splunk, the response of the Splunk query depends on query data**

| Duration | Good Traffic(%) | Exceptions /Bad Traffic(%) | CSAT(%) |
|---|---|---|---|
| FY19 Q3 | 98.316161 | 1.683839 | 54.7 |
| FY19 Q4 | 98.673637 | 1.326363 | 59.7 |
| FY20Q1 | 98.860992 | 1.139008 | 61.4 |

Alternatively, effort was made to make the relationship between Time taken to browse the page but it's creating many to many relationships between Time taken and Browser code. Also, tried to make relationship with user browser with browser code also, it's creating many to many relationships.

**Figure 5:** Comparison showing Good Vs Exception Vs CSAT Traffic.




**8.Conclusion**

After doing the multiple regression, Correlogram and Excel analysis of data we can conclude customer satisfaction is directly related with increase of good traffic.

Customer Satisfaction depends on below question which were answered by the Analysis:

- Did you accomplish the goal of your visit?
- How likely are you to recommend us to your friend or colleague?
- What can we do to improve your experience?
- What do you like the most about visiting our website?
- Are you going to return to our website?

**References (APA)**

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2009). Marketing Segmentation Through Machine Learning Models: An Approach Based on Customer Relationship Management and Customer Profitability Accounting. *Social Science Computer Review*, *27*(1), 96–117

QAS. (2006, September). The hidden costs of poor data management (International Research White Paper) . Dynamic Markets Commissioned by QAS.

Kim, Y. , Street, W.N. , Russell, G.J. , &Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithm. Management Science, 51, 264-276.

He, Z. , Xu, X. , Huang, J.Z. , & Deng, S. (2004). Mining class outliers: Concepts, algorithms and applications in CRM. Expert Systems with Applications, 27, 681-697.

Kumar, V. ,& Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century. Journal of Retailing, 80, 317-330.

Malmi, T. ,Raulas, M. , Gudergan, S. , &Sehm, J. (2004). An empirical study on customer profitability accounting, customer orientation and business unit performance (Draft paper). Helsinki School of Economics & Business Administration.

Van Raaij, V., Vernooij, M.J.A. & Van Triest, S. (2003). The implementation of customer profitability analysis: A case study. Industrial Marketing Management, 32, 573-583.

Winer, R.S. (2001). Customer relationship management: A framework, research directions, and the future (Draft Paper). University of California, Berkeley.

Wedel, M, & Kamakura, W.A. (2000). Market segmentation: Conceptual and methodological foundations. Boston: Kluwer Academic.

Murthy, S.K. (1998). Automatic construction of decision trees from data: A multidisciplinary survey. Data Mining and Knowledge Discovery, 2, 345-389.

Berger, P.D. ,& Nasr, N. (1998). Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing, 12, 17-30.

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning, Artificial Intelligence, 97, 245-271.

Rao, V.R. ,&Steckel, J.H. (1995). Selecting, evaluating and updating prospects in direct mail marketing. Journal of Direct Marketing, 9, 20-31.

Dibb, S., & Stern, P. (1995). Questioning the reliability of market segmentation techniques. Omega, 23, 625-636.

John, G.H. ,Kohavi, R. , &Pfleger, K. (1994, July). Irrelevant features and the subset selection problem. In W. W. Cohen & H. Hirsh (Eds.), Proceedings of the Eleventh International Conference on Machine Learning (pp. 121-129). San Francisco: Morgan Kaufmann.

Schmittlein, D.C. & Petersen, R.A. (1994). Customer base analysis: An industrial purchase process application. Marketing Science 13, 41-67.

Heath, D. ,Kasif, S. , &Salzberg, S. (1993, August-September). Learning oblique decision trees In R. Bajcsy (Ed.), Proceedings of the 13th International. Joint Conference on Artificial Intelligence (pp. 1002-1007). San Francisco: Morgan Kaufmann

Cooper, R. ,& Kaplan, R.S. (1988). Measure cost right: Make the right decision. Harvard Business Review, September-October, 96-103

Gath, I. ,&Geva, A.B. (1988). Unsupervised optimal fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7), 773-781.

Breiman, L, Friedman, J.H. ,Olshen, R.A. , & Stone, C.J. (1984). Classification and Regression Trees. New York: Chapman & Hall.

Green, P.E. ,& Wind, Y. (1975). New ways to measure consumers' judgments. Harvard Business Review, 53, 107-

Anderberg, M.R. (1973). Cluster analysis for applications. New York: Academic Press.

Bass, F.M., Tigert, D.G. & Lonsdale, R.T. (1968). Market segmentation: Group versus individual behavior. Journal of Marketing Research, 5, 264-270.