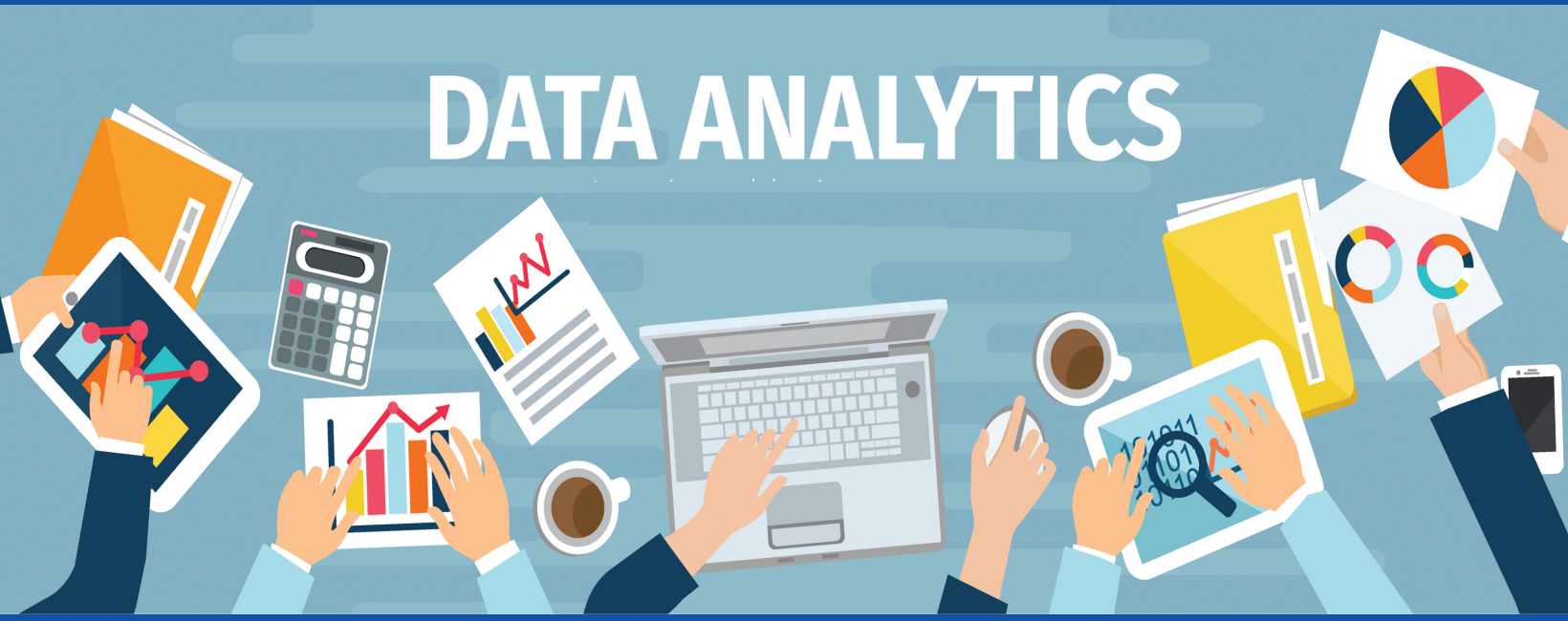
Volume 5 Issue 2

ISSN: 2321-1857

# International Journal of

# Business Analytics & Intelligence

October 2017




A Publication of Publishing India Group

# International Journal of Business Analytics & Intelligence

# Editorial Message

Greetings from team IJBAI!!!

It's indeed a matter of great joy for IJBAI fraternity when both the Editor-in-Chief and the Joint Editor-in-Chief of IJBAI were awarded as 'Analytics and Insight Leader of the Year' at a grand ceremony organized by KamiKaze at Vivanta by Taj, Bengaluru on 20th July, 2017. Secondly, we are delighted to have Professor U Dinesh Kumar of IIM Bangalore to have amongst us as a member of our coveted editorial board. He is professor of decision sciences and information systems, chairperson of executive PGPM, chairperson of decision sciences and information systems and chairperson of data centre and analytics lab. Last but not the least, we are honoured to have Prof. Banibrata Roy in our editorial board who is currently teaching in both University of Manitoba, Canada and Fairleigh Dickinson University, USA. Thus, with the series of ecstatic notes, it is indeed a great pleasure for us to publish the 2nd issue of Volume 5 in this year 2017.

At IJBAI, our vision is to disseminate the latest research and development of data science practice to our readers throughout its path of evolution. True to our vision, the journal becomes a platform for the exchange of the knowledge and insights on business analytics eco-system, analytical techniques and applications of analytics. In the present issue of IJBAI, we are delighted to bring out five application oriented domain focused analytics research papers along with two insight-stimulating perspectives from the stalwart of analytics academia, Prof. Arnab Laha of IIM Ahmedabad and the industry leader Ms. Madhumita Ghosh of IBM.

Prof. Arnab Laha's column "Analytically Yours", which is immensely popular among the readers, comprises his perspective on analytics applied to data streams. Ms. Madhumita Ghosh in her column on quick steps to deal in "data forest" provides a flavor to readers about the approach of data processing and management when the disruption of technology and business need to interpret 5Vs (Velocity, Volume, Variety, Veracity and Value) faster than firm's competitor can make or break a business. This requires outlining best practices to manage data governance, an exploration of data stewardship and details about common problems that firms experiences while instituting strong data management.

It is now no secret that customer retention is a top priority for almost all companies; acquiring new customers can be several times more expensive than retaining existing ones. Furthermore, gaining an understanding of the reasons customer's attrition and estimating the risk associated with individual customers are both powerful components of designing a data-driven retention strategy. Prof. Dash along with his student depicted how to identify the factors affecting customer attrition of trust accounts for a leading financial services company. In one hand, when focus is needed for customer attrition, on the other hand continuous inclusion of prospective customers by targeted marketing. We published a study by Prof. Prasad and Prof. Anjaneyulu which exhibits how to better support marketing decision makers in identifying better prospective customers by using Generalized Additive Models (GAM) and it's comparison over Logistic Regression. In the customer journey, experience plays a vital role. On that note, in the paper by Prof. Ghosh Dastidar, a solution for a restaurant recommendation system is presented which recommends a list of restaurants to the users based on their preference criteria. Customer experience improves through customer feedback. Professor Srivastava explores the social media strategies deployed by e-commerce companies by analyzing customer tweets through identification of polarities using naïve Bayes algorithm. We cannot ignore the fact that that logistics and inventory management are increasingly important to customer experiences. While most retailers understand the importance of providing a consistently positive customer experience, many struggles with legacy technology that fails to address customer experience needs, keeping the cost optimum. Professor Ajith Kumar along with his couple of students at XLRI analyzes replacement policy in a two-echelon supply chain using discrete-event simulation.

We wish to create this IJBAI, a leading repository of knowledge in analytics. A number of constructive steps, including the creation of the most scholastic editorial board, are taken to constantly improve the quality of the journal and thereby delight our esteemed readers. We are sure that our readers will enjoy and learn a lot from the present issue. Do let us know your wish, suggestions and views to enrich our journal. Therefore, it would be great to have valuable feedback from our learned readers about the enriched version of IJBAI. We would like to thank all the researchers and renowned data science practitioners who have honored us by selecting our journal to publish some of their research cases. At the end, we extend our heartfelt thanks to all our esteemed readers who continued to support us for the last five years.

Sincerely yours,
Madhumita Ghosh
Joint Editor-in-Chief
&
Tuhin Chattopadhyay
Editor-in-Chief
Dated: 1st August, 2017

# International Journal of Business Analytics and Intelligence

## Volume 5 Issue 2 October 2017

ISSN: 2321-1857





Journal is also available online at www.publishingindia.com

Article can be accessed online at http://www.publishingindia.com

# Analytically Yours:

# Analysis of Data Streams

**Arnab Kumar Laha***

One of the four Vs of Big data is 'velocity' which refers to the fact that in many applications, data is not static but continuously flows into the system (often at a very high rate). Such continuously flowing data is termed as Streaming data and is generated by various sources such as surveillance cameras, sensors in machines such as aircraft engines, tractors, vehicles and mobile phones, atmospheric systems, mass production systems, transactions such as those of a credit card system etc. In some applications such as credit card fraud detection, intrusion detection or preventive maintenance it is important that we are able to analyse the data stream in real time. There are two major challenges associated with analysis of data streams which are not present when dealing with static data. Firstly, it is not possible to work with the 'whole data' as the data keeps flowing into the system and secondly, the nature of the data changes over time- a phenomenon often referred to as 'concept drift'. Thus, analysis of streaming data requires techniques different from those used for static data.

Since the data flows continuously and often at a very high rate it necessitates the use of techniques that allow us to update the results quickly as new data is accumulated. Specifically, techniques that require use of the entire data available at each point of time or that require multiple passes over the entire data set are often not suitable for use with streaming data. Moreover, as concept drift is commonly present in streaming data we need to monitor the results (or output) quite closely to detect the occurrence of a change in the data generating system so that the model used can be updated. When concept drift is present in the streaming data, it is not even appropriate to use the entire historical data for building the model. Instead, researchers have suggested several alternative methods. Among these, the use of a 'data window' is the most popular. In this approach the model is built using a subset of data typically the most recent. Once a model is built, its predictive performance is tracked and when the model's performance deteriorates, it is re-built using the most recent window of data. The window size i.e. the number of recent observations to be included in the window, is chosen keeping in mind both the accuracy of the predictions as well as the computation time required to build the model. The second factor is important for real time applications.

Let us illustrate the above ideas using a simple example. Suppose we have streaming data about two related variables X (say, distance travelled by a passenger in a taxi) and Y (say, taxi fare) and we are interested in predicting Y based on X. We consider a simple linear regression model and compare two strategies: (A) build the model with the first 100 observations and use the same for predicting, and (B) build the model with the first 100 observations and then keep rebuilding the model with the latest 100 observations whenever 500 predictions are completed. We compare these strategies in two different scenarios: In scenario I, there is no concept drift while in scenario II, concept drift is present. Specifically in scenario I, the data generating mechanism $Y_i = 1 + 2X_i + \in_i$ is where $\in_i \sim N(\mu = 0, \sigma = 2)$, i = 1,…,3000 whereas in scenario II, the data generating mechanism is

$Y_i = 1 + 2X_i + \in_i$ where $\in_i \sim N(\mu = 0, \sigma = 2)$, $i = 1,\ldots,1000$,

$Y_i = 1 + 1.5X_i + \in_i$ where $\in_i \sim N(\mu = 0, \sigma = 2)$, $i = 1001,\ldots,2000$ and,

$Y_i = 1 + X_i + \in_i$ where $\in_i \sim N(\mu = 0, \sigma = 2)$, $i = 2001,\ldots,3000$.

We begin our discussion by comparing the performance of two strategies in scenario I. We use the Root Mean Square Prediction Error (RMSPE) as the measure of performance. It is computed as

* Indian Institute of Management, Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where $\hat{y}_i$ and $y_i$ are the predicted and actual value. We find the RMSPE values of the two strategies are very close to one another as would be expected in this case ($RMSPE_A$ = 1.98 and $RMSPE_B$ = 1.99). Thus, periodic rebuilding of the model does not yield any benefit in this case.

Now, let us examine how these two strategies perform in scenario II. We find that the RMSPE value of strategy A is far more than that of strategy B ($RMSPE_A$ = 6.95 and $MSPE_B$ = 2.35) indicating that strategy A performs quite poorly compared to strategy B. Thus, we find that the periodic rebuilding of the model used in strategy B had a positive impact in terms of improving the prediction quality. In general, periodic rebuilding of the model is a recommended strategy when dealing with streaming data where concept drift is suspected to be present.

A natural question that arises at this stage is how one knows that a rebuilding of the model is required. One way is to use the prediction errors. Intuitively we feel that if the prediction errors become larger than expected, then we should rebuild the model with current data. A widely used tool for handling such problems is the control chart introduced by W. Shewhart in the context of industrial quality control. We now discuss how a control chart can be used for monitoring the model output. For the purpose of this illustration, we restrict ourselves to scenario II. As before we build the model using the first 100 observations and obtain the residuals (residual = actual value - fitted value). These residuals are used for creating an individuals control chart for monitoring prediction errors. Since the average value of the residuals in linear regression is 0, we can use that as the centre line of the individuals control chart. The Upper Control Limit (UCL) in the individuals control chart is set to 3 s and the Lower Control Limit (LCL) is set to –3s where s is the standard deviation of the residuals. This individuals control chart can be used for monitoring the model performance as follows. Each prediction error is plotted on the control chart taking care that the same sequence as that of the arrivals of the observations is maintained. If a prediction error falls above the UCL or below the LCL we decide to rebuild the model. Fig. 1 gives the prediction errors plotted on the control chart which is created using the R package qcc. As can be seen from the chart that the prediction error of the rightmost observation is above the UCL indicating the need for rebuilding the model.

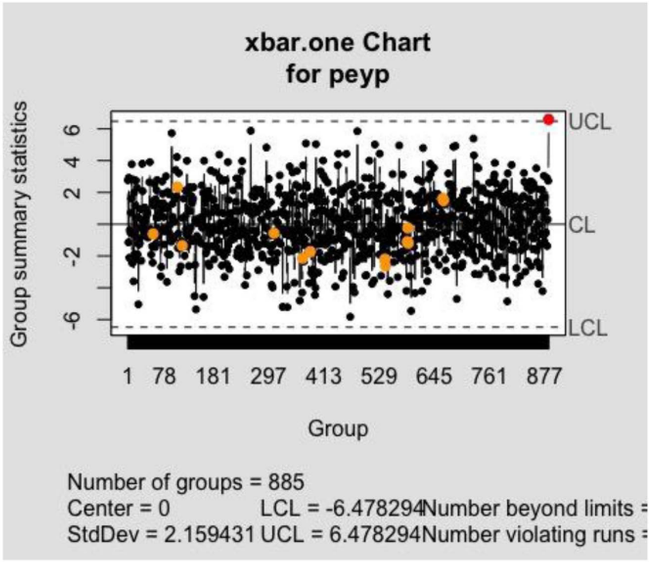



**Fig. 1: Prediction Errors Plotted on the Control Chart**

In summary, we can say that when dealing with streaming data with possibility of concept drift it is important for us to track the performance of the model and rebuild it as and when necessary. The individuals control chart for prediction errors can provide guidance regarding whether the current model is adequate or there is a need to rebuild the model with current data.

Article can be accessed online at http://www.publishingindia.com

# Quick Ready Steps to Deal in "Data" Forest

**Madhumita Ghosh***

Every organisation has some degree and magnitude of 'chaotic' culture, especially when it comes to data organisation and data management. Some of them breed chaos and unconsciously operate in chaos. Data management is designed to operate with structure. However, reality has always contained a dose of 'wonderland' as well. Data resettlement is typically the most neglected or forgotten component of any IT project. It is a process of migrating from one system to another, preferably without disrupting or disabling active business processes.

On some occasions, it is not easy to understand or probably viable to consider that the data migration is needed in the project and most of the times data migration is not seen as an item whose requirements need to be captured during analysis phase. That's why, migration related problems begin either during development phase or in testing phase when data need is identified or when the data from the old system refuses to fit properly into your new user interfaces or business rules despite transportation of the old data.

For a successful project, and to carry out an insightful analysis the need of data migration and its requirements must be identified early during beginning of analysis phase and further actions should be reflected to project plan accordingly, but how? Capturing the essentials of data nature is a life-saver activity which is the targeted idea of this work.

## Is Data Migration Really a Need?

Understanding the business needs and thereby translating those into analytical needs are utmost essential which should be identified during elicitation phase. If not possible during elicitation phase, the same can be done during analysis phase, but one needs to be careful and should identify the needs as earlier as possible.

One needs to enquire data related simple questions:

- Need the requirements mentioned from the systems be shut down? If yes, does this system keep any data? What type of information is it?
- Are the requirements mentioned from entities already used in company's systems? In order to use that data, need the new system's data model keep them as well? If yes, should this be kept in the same format or otherwise?

If either of the answer is yes, one will need a potential data migration. Moreover, it is advisable to get more details. Fig. 1 is a reference to remember the building blocks for data management

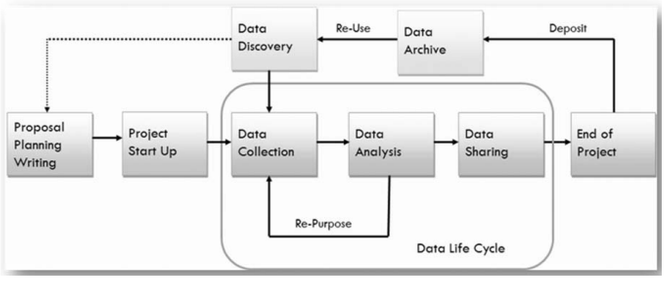



**Fig. 1: Building Blocks for Data Management**

## How to Capture Requirements and Define Business Rules?

If the group or department is determined that it needs a data migration, be careful and specify at least following details in the situation analysis.

### To Identify if Company has Multiple Data Sources for that Entities

If yes, determine the master and the fate of other systems' data. You will see how many systems you are dealing

* Practice Leader - Big Data & Advanced Analysis BA & Strategy - Global Business Services IBM, India.

with. It is important to know how and where the data is stored, backed up, and if it is archived.

## Data Profiling

Data profiling is a 'must have' activity to understand what are specs of the material the team/ department will be working on. They need to conduct profiling activities and classify the data, such as data with missing unique IDs or missing name or false data etc. Results of the profiling guides to design user interfaces and capture business rules.


**Fig. 2: The Cycle of Pre-Analytics Data Processing**

## Data Cleansing

According the data profiling, one might need cleansing activities for the final structure of the data. Also, it is required to clarify cleansing related activities on the project timeline to have a clearer understanding for the project timeline. If the answer is no, definitely providence of missing or dirty data to be checked.

## Data Structures

Data structures must be well understood if the project requires design of user interfaces or data forms. Consideration is must for the data structures while determining dynamic user interfaces or form context.

For example, old system may store the name and its salutation in one field by separating data by " – ". This means two information are stored in one field and one should consider such constraints while you are designing a new interface with multiple fields on screen validation and rules or determine ways to separate salutation and name effectively while moving the data as mentioned on section on data mapping.

## Providence of the Historical Data

It is needed to take into account if historical data will be migrated or not. It may affect user interfaces or business processes.

## Providence of the Missing or Dirty Data

After profiling, most probably it will be seen that some of the data is not clean or adequate to be used in further actions. For example, study is with sensible customer data and in that case unique identity number is mandatory but some records do not have identity numbers. It will cause problem to pinpoint the customer or it will be further problems if this information is mandatory to display customer on the screen. Even worse, if one has validations based on the identity number such as debt control or billing, the system will not be able to conduct such validations.

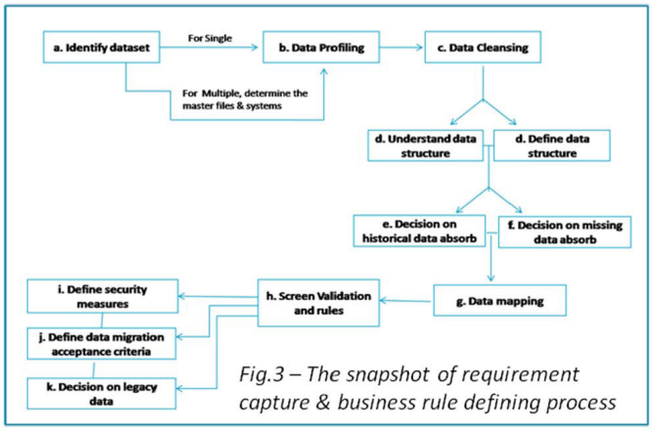



**Fig. 3: Snapshot of Requirement Capture and Business Rule Defining Process**

*Always check whether data ownership belongs to a specific business unit. If yes, let them decide to the fate of data.*

- It is advisable to decide whether the data is just enough to use or will it cause problems to conduct business processes. To migrate such data, sections on screen validations & rules and migration acceptance criteria will be highly important.
- Is it possible to clean or enhance problematic data somehow? If yes, determine the ways and related requirements.
- If one decides not to move such data, he/ she should always consult the respective business unit for possible further actions. They may need to find the customer and inform him/ her legally according his/ her account status or they may need another manual/ automation processes.

### Data Mapping

Data mapping is basically the activity of creating a map of the existing data model by matching each entity & field with the future data model. Each entity should be mapped correctly and in details to be able to move data successfully. The map is an essential item of *data migration strategy*. Based on the mapping, one can see the gap between legacies and future and one can use the information on screen validations and rules.

### Screen Validations and Rules

Results of earlier sections will provide clear information about user interface validations and potential need of new business processes.

#### Screen Validations

- To define entity specs such as type and length based on the profiling results, the information will guide to design potential data forms.
- To define the rules for the gap: If entities are not matching neatly, need to define UI standards and validation rules accordingly.

#### New Processes are Needed, in Case One Decided to Transport Problematic Data

- One may need to create new processes to correct such data. For example if the identity number is mandatory and your customer update process originally does not allow to update the number, in order to enable the correction of the customer information, one may need the define rules such as displaying identity number field editable.
- One may create new processes to alert the system or trigger different actions.

### Define Security & Security Measures Such as Encryption

If data needs to be migrated in encrypted form, general rules shall be set during analysis phase.

### Define Migration Acceptance Criteria

Define migration acceptance criteria such as data quality, migration duration etc. if they cause termination on your services.

### Define the Fate of the Legacy Data

Is it advisable to keep the data on the legacy systems? If yes, determine whether synchronisation is needed with the new system or not? How long the data should be kept on legacy system? What are maintenance rules?

## What Next?

Of course, one should plan the future analysis steps! All answers will guide for further activities which need to be added to the project plan in details. Identifying these steps at the beginning will prevent from future unexpected surprises and definitely will help to close the project on time. One is rehearsals and the other is testing scenarios.

## Conclusion

A core belief in the traditionalist viewpoint of IT project is that business analysis focuses on 'what' the client wants while design concentrates on 'how' to deliver the solution. As a business analysis practitioner, one can discover and describe a requirement that specifies the ability of a user to capture their state (or province) of residence. The main purpose of business analytics IT projects is to develop systems that turn large and often highly complex data sets

into meaningful information from which decisions can be made.

Putting together a comprehensive data migration plan now can help firms avoid serious downtime later, according to experts. Data migration is almost never done simply for the sake of data migration. It has to be tied back to something in the business, and IT managers need to generate their case around those business anchors.

Article can be accessed online at http://www.publishingindia.com

# Customer Attrition Analytics in Banking

**Mihir Dash*, Kajal Das****

**Abstract**

In an era of mature markets and intensive competitive pressure, more and more companies realise that their most precious asset is their existing customer base. This realisation has resulted in a rise in emphasis on customer relationship management, in order to retain customers. This is a major area on which banks need to concentrate. Banks tend to be reactive to customer attrition, and many times it is too late to retain a customer. Customer attrition needs to be minimised, and loyal customers need to be rewarded.

The objective of this study to identify the factors affecting customer attrition of trust accounts for a leading American financial services company. The company realised that its trust accounts were getting closed after a period of seven to twelve years. Initially, the company tried to identify the root cause using a small set of data, but they were unable to do so. This triggered the use of analytics to build a model to predict customer churn, and come up with strategies to retain customers. This was achieved by applying data mining techniques to the transactions history of the accounts that closed down as against those that remained active.

**Keywords:** Customer Attrition Analytics, Customer Relationship Management, Data Mining

## Introduction

Customer attrition is the phenomenon wherein a customer leaves a service provider. With the growing competition in the service sector, preventing customer attrition has become critical for sustainability, as it is well-established that retaining existing customers is more profitable than acquiring new customers (Jacob, 1994). This gives customer attrition analytics the challenging task of predicting which customers are likely to leave, and of subsequently designing and implementing retention programmes for these customers. Customer analytics has made many strides in marketing, employer desirability, and branding.

Several statistical techniques are commonly applied for customer attrition analytics, including classification & regression trees (Gray & Fan, 2008), logistic regression (Au, Chan, & Yao, 2003), artificial neural networks (Datta, Masand, Mani, & Li, 2001), survival analysis (Ma & Li, 1994), and several other techniques (Hadden, Tiwari, Roy, & Ruta, 2006). There are mixed results concerning the most appropriate technique; however, several studies support the logistic regression model. For instance, Mozer, Wolniewicz, Grimes, Johnson, and Kaushansky (2000) and Hwang, Jung, and Suh (2004) suggested that logistic regression predicted customer attrition better than decision trees and neural networks.

The case is about an American financial services company headquartered in Dallas, Texas. It has retail-banking operations in Arizona, California and Florida, with select business operations in several other U.S. states, as well as in Canada and Mexico. It is among the twenty-five largest U.S. financial holding companies, with $50 billion in total assets, $35 billion in total loans, and $40 billion in total deposits as of June 30, 2016. Its major operating units include the Business Bank, the Retail Bank, and Wealth Management. The case involves two of its trust accounts, viz. Investment Management Accounts and Revocable Living Trust accounts.

An Investment Management Account (IMA) is a flexible fund management arrangement with a financial institution that allows the customer to diversify their portfolio by gaining access to a wide range of financial instruments that span various asset types. Through an IMA account, customers can avail the wealth management services provided by fund managers under the auspices of the

* Professor & Head of Department, Management Science, School of Business, Alliance University, Karnataka, India. Email: mihirda@rediffmail.com
** Executive Student, School of Business, Alliance University, Karnataka, India. Email: daskajal@yahoo.com

bank. The fund managers implement investment strategies to meet the individual customers' financial goals. A Revocable Living Trust account provides financial protection for the holder in case of incapacity. The bank has several other forms of trust accounts, including Marital Trust accounts, which provide lifetime financial protection for spouses, Qualified Terminable Interest Trust accounts, which balance the financial interests of spouses and children in the event of death of the holder, Gifts-to-Minors Trust accounts, which set aside funds for young children, and so on.

Customers for banking and financial services expect the pleasing, intuitive digital experiences that have become routine in many aspects of daily life. However, instead of this personalised approach attuned to their individual preferences, traditional banking has been hampered by regulatory issues and complex processes. This in turn can make offering new products and services at the speed and scale of evolving customer demands difficult. As a result, consumers often face channel experiences that reflect the isolated, siloed information that institutions have about their customers.

Analytics has been used widely by banks to ensure they remain competitive in three areas of digital disruption: the customer interface, process digitalisation and data analytics. These include activities such as strategic, advisory and management consulting across analytics and information management, business intelligence and analytics; strategic, advisory and implementation services for next-generation technologies such as big data analytics, mobile and cloud business intelligence; enterprise performance management solutions spanning business strategy and enterprise metrics definitions; and end-to-end information management services including data management, data integration, data quality and data governance.

## Problem Statement

Acquiring new customers is usually harder and more expensive than retaining existing customers, who already know and trust the company, and in turn the company already knows so much about them. Retaining them might be as simple as making a phone call, providing free service for some components, or some other gesture of goodwill. The key is making full use of historical data of transaction, customer behaviour, and so on to understand what makes customers leave. Once this has been identified, the company can focus on trying to retain customers with highest value and having the highest risk of leaving.

The case concerns an American financial services company that has witnessed a 15% decline annually in revenue from its IMA trust accounts in the last four to five years. The bank realised that trust accounts were getting closed after operating for seven to twelve years, resulting in a loss in market share, declining revenue, and declining brokerage, while the cost of acquiring new customers was increasing as the bank had to spend a lot of its marketing efforts to acquire a new account; moreover, the number of accounts closed was higher than the number of new accounts acquired.

The objective of the study was to predict which IMA trust accounts were at risk of attrition. The initial hypothesis was that accounts were at risk of attrition if the sell transactions were greater than the buy transactions.

## Methodology

Predictive models relate the performance or behaviour of a typical unit in a system with some known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in the system will exhibit similar performance or behaviour, depending on its attributes or features.

The problem addressed by the study is a classification problem: the goal is to classify whether a trust account is likely to become inoperative. The study uses logistic regression and decision tree models for this purpose. Both of these techniques employ a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data.

The research was conducted in four phases; viz. source data integration, exploratory data analysis, model building and validation, and finally model scoring.

## Data Collection

The following data were collected as independent variables for the predictive model for customer attrition: account demographics, which gives details of the IMA and revocable account types; granular-level sell transaction details, which tell the number of sell transactions

(portfolio, bond, securities) by trust account; granular buy transaction details, i.e. details of securities or portfolio purchased; and Moody's/ Fed's macroeconomic data, which includes a variety of economy-wide phenomena such as as inflation, price levels, the rates of growth of national income/gross domestic product, and changes in unemployment, reflecting market sentiment. The details included under each category are presented below.

## Account Demographics

Table 1 presented the data for the IMA and revocable account types. When a trust opens an account, the demographic and investment details are captured in this table. The table captures the type of account, region-branch-date opened. The table also captures the time horizon for investment made (tenure), risk tolerance and anticipated market value to achieved after tenure, from an investment perspective. The attrition model uses the account demographic data, open/close indicator, and investment objective.

**Table 1: Account Demographic Variables**

| *Account Type* | *IMA or revocable account* |
|---|---|
| Date Opened | Account open date |
| Date Closed | Account closed date |
| Reason For Closing | Account closing reason |
| Risk Tolerance | Risk appetite of trust account along with horizon (closely related to investment strategy); typical values are : Balanced, Conservative Short-intermediate Conservative_Short_Intermediate Defensive_Growth_Intermediate Balanced_Intermediate Conservative_Intermediate |
| Investment Horizon | Total length of time that an investor expects to hold a security or a portfolio. |
| Investment Objective | The objective of the investment; typical values are: increase income , reduce tax liability, increase saving Minimum_Risk Moderate_No_Alt Maximum_Risk_No_Alt Moderately_Conservative_No_Alt Siadfm Camdfi Bonds_Only |

## Sell & Buy Transaction Tables

Table 2 captures buy/ sell activity done on behalf of the trust. Every buy/ sell transaction done by the company is recorded in this table along with security name purchased/ sold, amount of buy/ sell and the transaction fee for the company. The attrition model uses the aggregated buy/ sell amount for the last one month, three months and six months which signifies the activity levels of the trust account.

**Table 2: Sell & Buy Transactions Variables**

| *Transaction Date* | *Date of transaction executed (sell or buy)* |
|---|---|
| CUSIP | Portfolio ID |
| Security Name | Name of the securities to purchase or sell |
| Security Type | Typical values are: Shares, Bonds open-end fund, Index open-end fund |
| Reversal | |
| Transaction Amount | Sell amount/purchase amount |
| Units | Number of units |
| Transaction Description | Descriptions of sell & buy transactions |

## Fed's Macroeconomic Data

Fed macroeconomic data (Table 3) is generally used to understand the market. The data includes a variety of economy-wide phenomena such as price levels, inflation, national income, gross domestic product, growth rates, and changes in unemployment. The attrition model uses future oil price index, S&P 500 volatility index, consumer confidence index, and bankruptcies as external influencing factors.

**Table 3: Fed's Macroeconomic Variables**

| | |
|---|---|
| FPDPGDP.US | NIPA: Chain-Type Price Index - GDP, (Index 2009=100, SA) |
| FYHHMEDQ.US | Median Household Income: All Races, (USD, SA) |
| FGDP.US | NIPA: Gross Domestic Product, (Bil. USD, SAAR) |
| FRGT3Y.US | Interest Rates: 3-Year Constant Maturity Securities, (% p.a., NSA) |
| FRILIBOR6M.US | LIBOR Rates: 6-Month U.S. Dollar Deposits, (% p.a., NSA) |
| FDJMIIUSDDWCFTDQ.US | Dow Jones U.S. total stock market index, (USD - closing price index, NSA) |
| FCBC.US | Consumer Confidence Index, (Index 1985=100, SA) |
| FLBR.US | Household Survey: Unemployment Rate, (%, SA) |
| FSPVOL.US | S&P 500 Volatility, (30 day MA, NSA) |
| FRFED.US | Interest Rates: Federal Funds Rate, (% p.a., NSA) |
| FEXS$.US | NIPA: Exports of Services, (Bil. Ch. 2009 USD, SAAR) |
| FYPSAVQ.US | Income: Personal - Saving Rate, (%, SAAR) |
| FCPIU.US | CPI: Urban Consumer - All Items, (Index 1982-84=100, SA) |
| FCPWTI.US | Futures Price: NYMEX Light Sweet Crude Oil - Contract 1, (USD per bbl, NSA) |
| FAHETP.US | Avg. Hourly Earnings: Private - Total, (USD per hr., SA) |
| FHST1.US | Housing Starts: Single-Family, (Mil. #, SAAR) |
| FPPISP3500.US | PPI: Finished Goods less Food and Energy, (Index 1982=100, SA) |
| FBKPY.US | Bankruptcies: Personal - Total - Trailing 12 months, (#, SA) |
| FCPGASQ.US | NYMEX Natural Gas Futures Prices: Contract 1, (USD per MMBtu, NSA) |
| FZPAVN.US | NIPA: Profits after Tax - Dividends, (Bil. USD, SAAR) |
| FRFHLMCFM.US | FHLMC: 30-Year Commitment Rate - Fixed Rate - National, (%, NSA) |
| FRPRIME.US | Interest Rates: Bank Prime Rate, (% p.a., NSA) |
| FRILIBOR1M.US | LIBOR Rates: 1-Month U.S. Dollar Deposits, (% p.a., NSA) |
| FIFNIP$Q.US | Private Fixed Investment: Nonresidential - Intellectual Property Products, (Bil. Ch. 2009 USD, SAAR) |
| FABABC.US | ABA Delinquency Rate: Bank Card Loans - Open End, (% of Loans 30+ Days Past Due, SA) |
| FMOF14TQ.US | Mortgage Originations: 1-4 Family - Total, (Bil. USD, SAAR) |
| FMOF14PQ.US | Mortgage Originations: 1-4 Family - Purchase, (Bil. USD, SAAR) |

The dependent variables included group trust accounts that had been closed between seven to twelve years, by reasons for attrition/ closure. The specific objective was to identify trust accounts that had left and had large and/ or more frequent transactions in a given month or quarter. The goal was to build an attrition model to identify high value and/or high frequency trust accounts that have high probability to closure.

The data for the study was primry data on the demographics and the transactions for a sample of 3257 unique trust accounts of the bank, collected over the two-year period 2013-15. Of the 3257 unique accounts, there were 2513 unique accounts with sell transactions data and 2543 unique accounts with buy transactions data. Further, of the 3257 unique accounts, 1648 accounts closed down during the study period, and only 1573 accounts remained open at the end of the study period. The major reasons for closure were: customer has switched to a competitor; customer found the fees too high; customer already had established relationship with other banks; bank failed to meet customer's expectations; customer has taken advice of an external financial advisor; and several other reasons. In most cases, closure was associated with lack of communication by the bank of its updated policy and new offers; further, with respect to handling of complaints, there was a lapse on the part of the bank in understanding the customers' problems and delayed or incorrect redressal. Further, for the 1648 closed accounts, sell and buy transactions data was available for 1481 accounts. Of the 1481 closed accounts for which sell and buy transactiosn were available, 926 accounts had more sell transactions than buy transactions, 289 accounts had equal sell and buy transactions, and 266 accounts had more buy transactions than sell transactions.

## Model Development

Exploratory data analysis was performed to summarise the main characteristics of the data sets. In this phase, data from different sources was integrated. The data was mainly primary data; only the macroeconomic data was secondary data. The data was collected from the company via a data discovery workshop in order to understand the variables and data pattern for the model development. Additionally, complaint data and relationship manager interaction data were captured to identify the reasons for closing of accounts. The sample was divided randomly, with 70% allocated for the traning set and 30% for the validation/test set.

To get better results, both logistic regression and decision tree methods have been employed for this classification problem. The model was designed and developed in R.

## Variable Identification/Selection

The first filter used for selecting appropriate variables was the information values (IV) as a criterion to identify variables with high predictive power. Table 4 indicates several variables had very low IV.

**Table 4: Variables Dropped with Very Low Information Value**

| Variable | Information Value | Strength |
|---|---|---|
| Rt_CONSERVATIVE_INTERMEDIATE | 0 | Very weak |
| Rt_CONSERVATIVE_LONG | 0 | Very weak |
| Rt_CONSERVATIVE_NA | 0 | Very weak |
| Rt_CONSERVATIVE_SHORT | 0 | Very weak |
| Rt_CONSERVATIVE_SHORT_INTERMEDIATE | 0 | Very weak |
| Rt_D...BALANCED.LONG | 0 | Very weak |
| Rt_DEFENSIVE.GROWTH_INTERMEDIATE | 0 | Very weak |
| Rt_DEFENSIVE.GROWTH_LONG | 0 | Very weak |
| Rt_DEFENSIVE.GROWTH_NA | 0 | Very weak |
| Rt_DEFENSIVE.GROWTH_SHORT | 0 | Very weak |
| Rt_DEFENSIVE.GROWTH_SHORT.INTERM | 0 | Very weak |
| Rt_GROWTH_INTERMEDIATE | 0 | Very weak |
| Rt_GROWTH_LONG | 0 | Very weak |
| Rt_GROWTH_SHORT | 0 | Very weak |
| Rt_GROWTH_SHORT_INTERMEDIATE | 0 | Very weak |
| Rt_NA_INTERMEDIATE | 0 | Very weak |
| Rt_NA_LONG | 0 | Very weak |
| Rt_NA_SHORT_INTERMEDIATE | 0 | Very weak |
| Is_MODERATE | 0 | Very weak |
| Rt_BALANCED_INTERMEDIATE | 0 | Very weak |
| Rt_BALANCED_SHORT | 0 | Very weak |

| Variable | Information Value | Strength |
|---|---|---|
| Is_MAXIMUM_RISK_NO_ALT | 0 | Very weak |
| Is_MODERATELY_AGGRESSIVE | 0 | Very weak |
| Is_MODERATELY_AGGRESSIVE_NO_ALT | 0 | Very weak |
| Is_MODERATELY_CONSERVATIVE | 0 | Very weak |
| Is_MODERATELY_CONSERVATIVE_NO_ALT | 0 | Very weak |
| Is_MODERATE_NO_ALT | 0 | Very weak |
| Is_MUNICIPALS_ONLY | 0 | Very weak |
| Is_NOT_APPLICABLE | 0 | Very weak |
| Is_SIADFM | 0 | Very weak |
| Rt_.GROWTH_NA | 0 | Very weak |
| Rt_.NA_SHORT | 0 | Very weak |
| Rt_AGGRESSIVE.GROWTH_INTERMEDIATE | 0 | Very weak |
| Rt_AGGRESSIVE.GROWTH_LONG | 0 | Very weak |
| Rt_AGGRESSIVE.GROWTH_NA | 0 | Very weak |
| Rt_AGGRESSIVE.GROWTH_SHORT | 0 | Very weak |
| Rt_AGGRESSIVE.GROWTH_SHORT_INTERM | 0 | Very weak |
| Rt_BALANCED.NA | 0 | Very weak |
| Rt_BALANCED_SHORT_INTERMEDIATE | 0 | Very weak |

As observed from Table 4, several of the investment strategy variables had very low IV, so that they were dropped from the analysis.

Further, for continuous variables, the variance inflation factor (VIF) was examined, and variables with high VIF (i.e. VIF > 5.0) were dropped. In the case of categorical varaibles, correlation between the categorical variables was adjusted. The key variables (with and without macroeceonomic varaibles) selected after eliminating multicollinearity (i.e. dropping variables with high VIF) are presented in Table 5.

**Table 5: Variables Selected after Eliminating Multicollinearity**

| With Macro Economic data | |
|---|---|
| **Variable** | **VIF** |
| Futures_Price_NYMEX_Light_Sweet_Crude_Oil_diff | 4.29 |
| Count_tran_6months_buy | 3.43 |
| Natural_Gas_Futures_Prices_diff | 3.27 |
| Count_tran_latest_buy | 3.06 |
| ABA_Delinquency_Rate_Bank_Card_Loans_diff | 2.90 |
| Sum_tran_latest_sell | 2.50 |
| Sum_tran_3months_sell | 2.49 |
| Count_tran_6months_sell | 2.34 |
| Tenure_month | 2.23 |
| Mortgage_Originations_Total_diff | 2.09 |
| Count_tran_latest_sell | 2.00 |
| Bankruptcies_diff | 1.93 |
| Income_Personal_Saving_Rate_diff | 1.87 |
| Consumer_Confidence_Index_diff | 1.82 |
| SP_500_Volatility_diff | 1.48 |

| Without Macro Economic data | |
|---|---|
| **Variables** | **VIF** |
| Sum_tran_latest_sell | 2.50 |
| Sum_tran_latest_buy | 1.00 |
| Sum_tran_3months_sell | 2.48 |
| Count_tran_latest_buy | 4.38 |
| Count_tran_latest_sell | 1.91 |
| Count_tran_3months_buy | 4.72 |
| Count_tran_6months_sell | 2.11 |
| Tenure_month | 1.06 |

With the inclusion of the macroeconomic data, the following variables were found to have low multicollinearity: Last Six Months Buy Transaction count, Future Prices of Crude Oil, Future Prices of Natural Gas, Delinquency rates of card loans, Mortgage Originations, Bankruptcies, S&P 500 Volatility index, and Consumer Confidence Index.

The last filter was to drop insignificant varables. Further, the results with and without macroeconomic data were compared, to see if there was any improvement in accuracy. The final set of variables (with and without macroeceonomic varaibles) selected after eliminating insignificant variables are presented in Table 6.

**Table 6: Final Set of Variables**

P Values _ without Macro Data

| Coefficients: | Pr(>\|z\|) | Significance level |
|---|---|---|
| (Intercept) | 2.71E-03 | ** |
| Sell_more_than_buy6_flag1 | 4.38E-16 | *** |
| Count_sell_more_equal_buy_latest_flag1 | 2.08E-02 | * |
| Count_sell_more_equal_buy6_flag1 | 1.76E-13 | *** |
| Sum_tran_3months_sell | 2.78E-05 | *** |
| Count_tran_latest_buy | <0.0000000000000002 | *** |
| Count_tran_3months_buy | <0.0000000000000002 | *** |
| Count_tran_6months_sell | 2.10E-10 | *** |
| Rt_.Growth_na1 | 3.68E-05 | *** |
| Rt_balanced_intermediate1 | 0.014288 | * |
| Rt_aggressive.Growth_intermediate1 | 0.005125 | ** |
| Rt_balanced.Na1 | 0.003631 | ** |
| Rt_defensive.Growth_short1 | 0.048882 | * |
| Rt_growth_intermediate1 | 0.002842 | ** |
| Is_.Minimum_risk1 | 0.010329 | * |
| Is_aggressive1 | 0.031026 | * |
| Is_bonds_only1 | 0.000394 | *** |
| Is_municipals_only1 | 0.00324 | ** |
| Is_municipals_only1 | 0.00324 | ** |

P Values _ with Macro Data

| Coefficients: | Pr(>\|z\|) | Significance Level |
|---|---|---|
| (Intercept) | 3.52E-06 | *** |
| Futures_price_nymex_light_sweet_crude_oil_diff | <0.0000000000000002 | *** |
| Count_tran_6months_buy | <0.0000000000000002 | *** |
| Natural_gas_futures_prices_diff | 1.34E-11 | *** |
| Count_tran_latest_buy | <0.0000000000000002 | *** |
| Sum_tran_3months_sell | 0.003177 | ** |
| Count_tran_6months_sell | 0.006253 | ** |
| Tenure_month | 4.34E-09 | *** |
| Mortgage_originations_total_diff | <0.0000000000000002 | *** |
| Count_tran_latest_sell | 0.046732 | * |
| Bankruptcies_diff | <0.0000000000000002 | *** |
| Sp_500_volatility_diff | 4.91E-05 | *** |
| Sell_more_than_buy6_flag | 1.44E-12 | *** |
| Count_sell_more_equal_buy_latest_flag | 1.21E-05 | *** |
| Count_sell_more_equal_buy6_flag | 1.22E-11 | *** |
| Rt_growth_na | 0.034466 | * |
| Rt_balanced_na | 0.003846 | ** |
| Rt_defensive_growth_short | 0.014521 | * |
| Rt_growth_intermediate | 0.036489 | * |

The following variables were found to be significant predictors after including macroeconomic data: Last Six Months Buy Transaction Count, Last Month Sell Transaction Count, Tenure, Futures Crude Oil, Futures Natural Gas, Mortgage Originations, Bankruptcies, and S&P 500 Volatility Index.

The models were compared on the basis of sensitivity, i.e. the proportion of actual positive cases (attrition customers) that were correctly identified, and specificity, i.e. the proportion of actual negative cases (non-attrition customers) that were correctly identified. The model results are summaried in the tables below. To check for robustness, the models were compared with decision tree models using the same underlying variables. The classification results for the test data set are presented in Table 7.

**Table 7: Classification Results for the Test Data**

Logistic Model – Without Macro data

| | |
|---|---|
| Sensitivity | 0.84 |
| Specificity | 0.69 |
| Accuracy of the Model | 0.77 |

Logistic Model – With Macro data

| | |
|---|---|
| Sensitivity | 0.83 |
| Specificity | 0.82 |
| Accuracy of the Model | 0.83 |

Key Insights:

- Accuracy of the Logistic Model has been increased by 6% after including Macro Data
- Sensitivity is decreased by 1 % and specificity increased by 13 % with Macro Data

Decision Tree – Without Macro data

| | |
|---|---|
| Sensitivity | 0.85 |
| Specificity | 0.78 |
| Accuracy of the Model | 0.82 |

Decision Tree – With Macro data

| | |
|---|---|
| Sensitivity | 0.91 |
| Specificity | 0.92 |
| Accuracy of the Model | 0.92 |

Key Insights:

- Accuracy of the Dtree Model has been increased by 10% after including Macro Data
- Sensitivity and specificity increased by 6 % and 14% respectively with Macro Data.

The classification results show that the decision teee model performs better than the logistic regression model in predicting both customer attrition and retention, particularly with the macroeconomic data. The final model was found to be quite accurate, i.e. 91% sensitivity and 92% specificity.

## Conclusion

Predictive analytics can be a powerful tool for customer retention. There are many predictive techniques that can help reduce customer attrition. In fact, this approach can also be used to predict other aspects of accounts behaviour, such as when a trust account will sell the portfolio, when a trust account will buy the portfolio, whether a customer is committing a financial crime, and so on. Banks and financial institutions can also use predictive techniques to improve marketing strategies for acquiring new customers.

In the current case, data mining was perfomed on past accounts, and it was observed that closed accounts have larger sell transactions than buy transactions for the last three to six months. By introducing macroeconomic data from the Fed, the final model was able to identify six/seven variables that were key predictors of account closure, with high accuracy in the model, viz. 92%.

The results of the study suggest that decision tree methods amy be more appropriate than logistic regression methods in customer attrition analysis. Thus finding is contrary to that of Mozer *et al.* (2000) and Hwang *et al*. (2004), who suggested that logistic regression was better than the decision tree model for predicting customer attrition. In the current case, decision tree model is perhaps more appropriate, as it is the event that sell transactions exceed buy transactions that predicts attrition, rather than the extent of the difference. It would be interesting to review the results of earlier studies with this additional insight.

The results of the study also suggest that macroeconomic variables play a significant role in customer attrition. Macroeconomic indicators clearly provide a stimulus for investment behaviour. The results of the study suggest that the revelant macroeconomic indicators for customer attrition are oil prices, natural gas prices, bankruptcies, and stock market volatility. The model suggests that there are threshold levels for each of these variables beyond which investors are more likely to exit the market.

The study contributes to the literature of customer attrition analytics by indicating that decision tree models may be more appropriate in some contexts of customer attrition, and by indicating that macroeconomic variables may play a signficant role in customer attrition modelling. There is great scope for extending the study in other customer attrition contexts.

## References

Au, W., Chan, C. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, 7, 532-545.

Datta, P., Masand, B., Mani, D. R., & Li, B. (2001). Automated cellular modelling and prediction on a large scale. *Issues on the Application of Data Mining*, 485-502.

Gray, J. B., & Fan, G. (2008). Classification tree analysis using TARGET. *Computational Statistics & Data Analysis,* 52, 1362-1372.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Churn prediction using complaints data. *Transactions on Engineering, Computing and Technology, Enformatika* 13, 158-163.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunications industry. *Expert Systems with Applications*, 26, 181-188.

Jacob, R. (1994). Why some customers are more equal than others. *Fortune*, 149-154.

Ma, G., & Li, S. (1994). Applications of the survival analysis techniques in modelling customer retention. *Workbook for the 5th Advanced Research Techniques Forum, American Marketing Association*.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Transactions on Neural Networks* 11, 690-696.

Article can be accessed online at http://www.publishingindia.com

# Application of Generalised Additive Logistic Model for Targeted Marketing

**K. V. N. K. Prasad*, G.V.S.R. Anjaneyulu****

**Abstract**

This study focuses on how to support marketing decision makers better in identifying better prospective customers by using generalised additive models (GAMs). Compared to logistic regression, GAM relaxes the linearity constraint which allows for complex non-linear fits to the data. In this paper, we examine how GAM-based logistic models perform compared to traditional logistic regression model and also provide some implications.

**Keywords:** Additive Logistics Model, Targeted Marketing

## Introduction

Last few years have seen an enormous emphasis on developing non-linear modeling approaches to deal with the problem of rapidly increasing variance for higher dimensional modeling problems because non-linear modeling approach provides information about the relationship between the dependent and independent variable that are not quite revealed by the standard modeling approaches. Secondly, a wide variety of model specifications can be tested so that model misspecification can be largely avoided (Hastie & Tibshirani, 1990).

Stone (1985) proposed the concept of additive models; these models estimate an additive approximation to the multivariate regression; in additive models each term is estimated by smoother – which helps in dealing with curse of dimensionality. Secondly, each estimate for each individual predictor in the model will explain how the dependent variable change in accordance with the corresponding individual predictors.

The extension of additive models across a wide range of distribution families (exponential family) by Hastie and Tibshirani (1990) called generalised additive models (GAMs). In GAM the mean of the dependent variable depends on an additive independent variable through a non-linear link function with a flexibility of response probability distribution being any member from the exponential family of distributions.

Generalised additive model proposed by Hastie and Tibshirani (1990) is a class of generalised linear model with a linear predictor involving a sum of smooth functions of covariates. The generalised additive model replaces $\sum_{j=1}^{p}\beta_j x_j$ (linear predictor) in generalised linear model architecture with $\sum_{j=1}^{p} f_j x_j$ (additive predictors) where $f_{j's}$ are unspecified non-parametric function. The generic form of generalised additive model can be represented as follows

$$\mu(x) = E\,[Y|X_1\,X_2\,....\,X_p\,] = (X_1\,)+(X_2\,).....(X_p\,) + \xi\,;$$

$$E\,[Y|X_1\,X_2\,....\,X_p\,] = + \xi\,;$$

where $y^{\sim}$ some exponential family of distributions

## Generalised Additive Logistic Model

The generic form of the logistic model in generalised linear model is as follows:

$$\text{Let } y = \begin{cases} 1 \text{ with probability } p(x) \\ 0 \text{ with probability } 1-p(x) \end{cases}$$

where $x = (X_1\,X_2\,....\,X_p\,)$ be a vector of covariates

then

$$\text{Logit } (p(x)) = \log\frac{p(X)}{1-p(X)}\,\beta_0\sum_{j=1}^{p}\beta_j x_j$$

and $$p(x) = \frac{Exp\,(\beta_0 + \sum_{j=1}^{p}\beta_j\left(x_{jj}\right)}{1+Exp\,(\beta_0 + \sum_{j=1}^{p}\beta_j\left(x_{jj}\right)}$$

* Department of Statistics, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. Email: kota.prasad.krishna@gmail.com
** Department of Statistics, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

The generalised additive logistic model assumes the functional form

$$\log \frac{p(Y \mid xi1,\ldots\ldots xip)}{1-p(Y \mid xi1,\ldots\ldots xip)} \beta_0 f_1(x_{i1}) + \ldots. + f_p(x_{ip})$$

$$= + f_1(x_{i1}) + \ldots. + f_p(x_{ip})$$

and $p(x) = \frac{Exp\,(f_0 + \sum_{j=1}^{p} f_j(x_{ij})}{1 + Exp\,(f_0 + \sum_{j=1}^{p} f_j(x_{ij})}$

$$\eta(x) = \log \frac{p(x)}{1-p(x)}$$

where $\eta$ is a function of $p$ variables.

Assume that $\eta(x) + \xi$ $Y =$ is an initial estimate of $\eta(x)$, the adjusted dependent variable is

$$Z_i = \eta_i + (y_i + \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

We fit additive model to $z_i$ 's, treating it as the response variable Y in E(Y|X) =+ $\alpha + \sum_{J=1}^{p} f_j x_j \xi$. The function $f_{\hat{j}} x_j$ is estimated by smoothing (where $f_{\hat{j}}$ are smooth functions of).

## Fitting Generalised Additive Models

The estimation of generalised additive models (GAMs) is a bi-folded iterative process. In each step, local scoring and back fitting algorithms are used until the convergence criteria for Backfitting algorithm is satisfied. If change in deviance of the estimates is below certain threshold, the algorithm stops, else based on estimates obtained from the Backfitting algorithm new weights are computed and iterative process is continued until the deviance of the estimates is below certain threshold.

### Backfitting and Local Scoring Algorithm

The Backfitting algorithm is a generic algorithm that can fit an additive model using any regression mechanism. The principal ideal behind Backfitting algorithm is to find the $j^{th}$ smoother of the $(k+1)^{th}$ iteration by smoothing the partial residuals defined by

$$R_j = Y - f_0 - \sum_{k \neq j} f_k(x_k)$$

The partial residuals remove the effects of all the other variables from y, thus they can be used to model the individual effect against $x_j$. The iterative mechanism in Backfitting algorithm constructs a smooth curve f(x) that summarises the dependency of y on x until the change in each effect is sufficiently small. This is achieved by starting with initial function and looping iteration cycle through partial residuals, fitting the individual smoothing components to its partial residuals. Hastie and Tibshirani (1990) proved that with many smoothers, the residual sum of squares (RSS) will never increase at any step; this implies that there are no convergence issues with the algorithm.

The general local scoring algorithm is an extension to the Backfitting Algorithm and is also an iterative algorithm. Since the linear predictors are replaced by additive predictors in generalised additive models, thus the Fisher scoring procedure is replaced by the local scoring algorithm as the predictions for the adjusted dependent variable are localised by nonparametric smoothers.

## Smoothing and Methods of Choosing Smoothing Parameter

Each smoother specified in generalised additive model has a single smoothing parameter. Smoothing is a mechanism of summarising the trend (non-parametric) in dependent variable Y as a function of one or more independent variables $X_1$, $X_2$, …. $X_p$. Since smoothing is just a summarisation of trend, it will not assume any rigid form of dependency/ relationship between Y and $X_1$, $X_2$, …. $X_p$. There are two methods for choosing smoothing parameters:

**(a) Cross-Validation:** In this method a data point ($x_i$, $y_i$) is left out at a time as testing set and the smoother is estimated at xi based on remaining n-1 points to minimise the sum of those squared residuals.

**(b) Generalised Cross-Validation:** It is a weighted cross-validation method which was proposed by Craven and Wahba (1979). The generalised cross-validation method provides an alternative and convenient approximation to the leave-one-out cross-validation with lesser computing cost. Minimising the generalised cross-validation criterion often yields a similar smoothing parameter to that obtained by the leave-one-out cross-validation.

## Introduction to Logistic Regression

Consider the following simple linear regression setting with 'r' predictor and binary response variable

$$y_i = \beta_0 + \beta_1 x_i + \ldots\ldots + \beta_r x_r + \int_i, i = 1, 2, \ldots . n$$

where $y_i$ is the binary response variable, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ ), and are independent.

Let $P_i$ denote the probability that $y_i = 1$ and $x_i = x$

$$P_i = P(Y_i = 1 | X_i = X) = \frac{1}{\left(1 + e^{-z}\right)}$$

where $Z = \beta_0 + \beta_1 x_i + \ldots\ldots + \beta_r x_r$

Or

$$\text{Logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_r x_r$$

The above equitation is called logistic regression, a statistical method in which we model the logit (p) in terms of explanatory variables that are available to modeler. It is non-linear in the parameters,……. The response probabilities are modeled by logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression.

## Maximum Likelihood Method of Estimation

In maximum likelihood estimation, we search over all possible sets of parameter values for a specified model to find the set of values for which the observed sample was most likely. That is, we find the set of parameter values that, given a model, were most likely to have given us the data that we have in hand.

## Model Performance Measures

The traditional model fit evaluation based on AIC and BIC can't be used for evaluation as - the AIC for generalised additive model is computed as deviance + 2p, where p are the model degrees of freedom; whereas the AIC for generalised linear models is computed as 2LL + 2p, where LL is the log likelihood of the fitted model and p are the model degrees of freedom.

Thus to evaluate the performance of the model we rely on the significance of the independent variables in both models, along with significance of the smoothing terms/ effects in the generalised additive models along with the following measures that are usually used to evaluate the performance of the model.

## Kolmogorov-Smirnov (KS)

This measures the maximum vertical separation (deviation) between the cumulative distributions of goods and bads and is defined as follows

$$KS = MAX\ |F_G^{(s)} - F_B^{(s)}|$$

The higher the KS value, the better is the model's ability for separation.

## Accuracy Ratio

Typically, the assessment of a marketing response model is evaluated by discriminatory power is commonly measured by the CAP or Lorenz curve (a variant of the ROC curve) and the accuracy ratio AR derived from this curve. The difference in comparison with the ROC curve consists in plotting the cumulative distribution function of all scores against that of the default score (instead of plotting the cumulative distribution functions of the non-default and default scores against each other). The accuracy ratio calculated from the CAP curve is however linearly related to the area under curve AUC of the ROC curve:

$$AR = 2AUC - 1$$

## Data

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The classification goal is to predict if the client will subscribe a term deposit. The dataset consist of 45211 instances.

## Basic Statistics and Correlation

The basic statistics for the numerical variables used in the model are reported below.

**Table 1: Statistics for Numerical Variables used in the Model**

| Variable | N | N Miss | Minimum | Maximum | Variance | Std Dev |
|---|---|---|---|---|---|---|
| age | 45211 | 0 | 18 | 95 | 112.76 | 10.62 |
| balance | 45211 | 0 | -8019 | 102127 | 9270598.95 | 3044.77 |
| day | 45211 | 0 | 1 | 31 | 69.26 | 8.32 |
| duration | 45211 | 0 | 0 | 4918 | 66320.57 | 257.53 |
| campaign | 45211 | 0 | 1 | 63 | 9.60 | 3.10 |
| pdays | 45211 | 0 | -1 | 871 | 10025.77 | 100.13 |
| previous | 45211 | 0 | 0 | 275 | 5.31 | 2.30 |
| target | 45211 | 0 | 0 | 1 | 0.10 | 0.32 |
| housing_n | 45211 | 0 | 0 | 1 | 0.25 | 0.50 |
| loan_n | 45211 | 0 | 0 | 1 | 0.13 | 0.37 |
| default_n | 45211 | 0 | 0 | 1 | 0.02 | 0.13 |
| job_n | 45211 | 0 | 1 | 2 | 0.25 | 0.50 |
| marital_n | 45211 | 0 | 1 | 2 | 0.24 | 0.49 |
| education_n | 45211 | 0 | 1 | 2 | 0.22 | 0.47 |
| contact_n | 45211 | 0 | 1 | 3 | 0.81 | 0.90 |
| poutcome_n | 45211 | 0 | 1 | 2 | 0.03 | 0.18 |
| _n_ variables are characterstic variables that are binned | | | | | | |

The following correlation analysis among the numeric variables indicates that there is no problem of co-linearity in the modeling dataset.

**Table 2: Correlation Analysis among the Numeric Variables**

| | age | balance | day | duration | campaign | pdays | previous | target | housing_n | loan_n | default_n | job_n | marital_n | education_n | contact_n | poutcome_n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1 | 0.098 | -0.009 | -0.005 | 0.005 | -0.024 | 0.001 | 0.025 | -0.186 | -0.016 | -0.018 | 0.102 | 0.286 | -0.050 | 0.026 | 0.036 |
| balance | | 1.000 | 0.005 | 0.022 | -0.015 | 0.003 | 0.017 | 0.053 | -0.069 | -0.084 | -0.067 | -0.038 | 0.026 | 0.086 | -0.027 | 0.035 |
| day | | | 1.000 | -0.030 | 0.162 | -0.093 | -0.052 | -0.028 | -0.028 | 0.011 | 0.009 | -0.031 | 0.007 | 0.021 | -0.028 | -0.030 |
| duration | | | | 1.000 | -0.085 | -0.002 | 0.001 | 0.395 | 0.005 | -0.012 | -0.010 | 0.015 | -0.023 | 0.001 | -0.021 | 0.042 |
| campaign | | | | | 1.000 | -0.089 | -0.033 | -0.073 | -0.024 | 0.010 | 0.017 | -0.013 | 0.031 | 0.015 | 0.020 | -0.057 |
| pdays | | | | | | 1.000 | 0.455 | 0.104 | 0.124 | -0.023 | -0.030 | 0.009 | -0.028 | -0.010 | -0.245 | 0.229 |
| previous | | | | | | | 1.000 | 0.093 | 0.037 | -0.011 | -0.018 | -0.019 | -0.013 | 0.019 | -0.148 | 0.201 |
| rand | | | | | | | | -0.005 | -0.001 | 0.000 | 0.005 | -0.013 | -0.007 | 0.009 | -0.003 | 0.000 |
| target | | | | | | | | 1.000 | -0.139 | -0.068 | -0.022 | -0.031 | -0.060 | 0.069 | -0.148 | 0.307 |
| housing_n | | | | | | | | | 1.000 | 0.041 | -0.006 | 0.072 | 0.018 | -0.116 | 0.188 | -0.091 |
| loan_n | | | | | | | | | | 1.000 | 0.077 | 0.021 | 0.037 | -0.065 | -0.011 | -0.054 |
| default_n | | | | | | | | | | | 1.000 | 0.008 | -0.014 | -0.015 | 0.015 | -0.023 |
| job_n | | | | | | | | | | | | 1.000 | 0.111 | -0.362 | 0.125 | -0.029 |
| marital_n | | | | | | | | | | | | | 1.000 | -0.089 | 0.033 | -0.018 |
| education_n | | | | | | | | | | | | | | 1.000 | -0.118 | 0.051 |
| contact_n | | | | | | | | | | | | | | | 1.000 | -0.114 |
| poutcome_n | | | | | | | | | | | | | | | | 1.000 |
| _n_ indicates characterstic variables that are binned | | | | | | | | | | | | | | | | |

## Training Vs Validation

The modeling data was split into 50/50 by using simple random sampling; the first 50% of the data was used to train the data and the remaining 50% of the data is used to validate the model.

**Table 3: Training vs Validation**

| Training Vs Validation: | | | | | | |
|---|---|---|---|---|---|---|
| Target | Frequency | Cumulative Frequency | Train Response rate | Frequency | Cumulative Frequency | Test Response rate |
| 0 | 19,920 | 19,920 | 11.88% | 20,002 | 20,002 | 11.52% |
| 1 | 2,685 | 22,605 | | 2,604 | 22,606 | |

## Logistic Vs Generalised Additive Logistic Model comparisons

**Table 4: Logistic vs Generalised Additive Logistic Model Comparisons**

| Logistic Regression | | | | | | |
|---|---|---|---|---|---|---|
| Analysis of Maximum Likelihood Estimates | | | | | | |
| Parameter | Categories | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | | 1 | -2.9686 | 0.1258 | 556.4165 | <.0001 |
| balance | | 1 | 0.000023 | 6.86E-06 | 11.5676 | 0.0007 |
| duration | | 1 | 0.00413 | 0.000088 | 2176.839 | <.0001 |
| age | | 1 | 0.00244 | 0.00234 | 1.0899 | 0.2965 |
| housing | 01: Yes | 1 | 0.3764 | 0.0266 | 200.9022 | <.0001 |
| | 02: No | | Ref | Ref | Ref | Ref |
| Loan | 01: Yes | 1 | 0.319 | 0.041 | 60.5037 | <.0001 |
| | 02: No | | Ref | Ref | Ref | Ref |
| poutcome_n | 01 : 'unknown','other','failure' | 1 | -1.2558 | 0.0447 | 787.6218 | <.0001 |
| | 01: 'success' | | Ref | Ref | Ref | Ref |
| marital_n | 01: 'divorced','single' | 1 | 0.1951 | 0.0263 | 55.0716 | <.0001 |
| | 02: Married & others | | Ref | Ref | Ref | Ref |
| education_n | 01: 'primary', 'secondary' | 1 | -0.0954 | 0.0257 | 13.7597 | 0.0002 |
| | 02: 'tertiary', 'others' | | Ref | Ref | Ref | Ref |
| contact_n | 01:'cellula' | 1 | 0.4875 | 0.0452 | 116.2995 | <.0001 |
| | 02:'telepho' | 1 | 0.281 | 0.0716 | 15.3937 | <.0001 |
| | 02:'Others' | | Ref | Ref | Ref | Ref |
| previous | | 1 | 0.0795 | 0.0104 | 57.8922 | <.0001 |

| Generalized Additive Logistic Regression | | | | | |
|---|---|---|---|---|---|
| Regression Model Analysis | | | | | |
| Parameter | Categories | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | | -2.8128 | 0.19068 | -14.75 | <.0001 |
| Linear(balance) | | -0.00051936 | 1.95E-05 | -26.64 | <.0001 |
| duration | | 0.00447 | 9.67E-05 | 46.22 | <.0001 |
| Linear(age) | | 0.00415 | 0.0024 | 1.73 | 0.0834 |
| housing | 01: Yes | 0.73103 | 0.0571 | 12.8 | <.0001 |
| | 02: No | Ref | Ref | Ref | Ref |
| Loan | 01: Yes | 0.57761 | 0.08671 | 6.66 | <.0001 |
| | 02: No | Ref | Ref | Ref | Ref |
| poutcome_n | 01 : 'unknown','other','failure' | -2.47527 | 0.09721 | -25.46 | <.0001 |
| | 01: 'success' | Ref | Ref | Ref | Ref |
| marital_n | 01: 'divorced','single' | 0.27462 | 0.05666 | 4.85 | <.0001 |
| | 02: Married & others | Ref | Ref | Ref | Ref |
| education_n | 01: 'primary', 'secondary' | -0.20967 | 0.05507 | -3.81 | 0.0001 |
| | 02: 'tertiary', 'others' | Ref | Ref | Ref | Ref |
| contact_n | 01:'cellula' | 1.28846 | 0.08539 | 15.09 | <.0001 |
| | 02:'telepho' | 0.83111 | 0.138 | 6.02 | <.0001 |
| | 02:'Others' | Ref | Ref | Ref | Ref |
| previous | | 0.09234 | 0.01117 | 8.27 | <.0001 |

From Table 4, we can observe that the estimate for age is highly insignificant, followed by balance in logistic regression, but in GAM the linear component of balance and age are statistically significant at 5% level of significance.

Table 5 provides a summary of significance of the smoothing parameters used in the model.

**Table 5: Summary of Significance of the Smoothing Parameters**

| Smoothing Model Analysis | | | |
|---|---|---|---|
| Approximate Analysis of Deviance | | | |
| Source | DF | Chi-Square | Pr > ChiSq |
| Spline(balance) | 352.1504 | 18801.4343 | <.0001 |
| Spline(age) | 12.43418 | 259.5807 | <.0001 |

For each smoothing effect in the model, Table 5 gives a $\chi^2$ test comparing the deviance between the full model and the model without the nonparametric component of this variable. The analysis of deviance results indicates that the non-parametric effect of age and balance are highly significant.

## Performance Analysis

**Table 6: Performance Results of Generalised Additive Models - I**

| Interval | Minmum Score | Maximum Score | Total Accounts | Number of goods | Number of bads | Good Rate | % goods | %bads | Cummulative goods : G(i) | Cummulative bads B(i) | KS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.60130 | 0.73110 | 2,260 | 1,232 | 1,028 | 54.51% | 45.88% | 5.16% | 45.88% | 5.16% | 40.72% |
| 2 | 0.53990 | 0.60130 | 2,261 | 706 | 1,555 | 31.23% | 26.29% | 7.81% | 72.18% | 12.97% | 59.21% |
| 3 | 0.52250 | 0.53990 | 2,260 | 308 | 1,952 | 13.63% | 11.47% | 9.80% | 83.65% | 22.77% | 60.88% |
| 4 | 0.51490 | 0.52250 | 2,261 | 166 | 2,095 | 7.34% | 6.18% | 10.52% | 89.83% | 33.28% | 56.55% |
| 5 | 0.51040 | 0.51490 | 2,260 | 75 | 2,185 | 3.32% | 2.79% | 10.97% | 92.63% | 44.25% | 48.37% |
| 6 | 0.50730 | 0.51040 | 2,261 | 48 | 2,213 | 2.12% | 1.79% | 11.11% | 94.41% | 55.36% | 39.05% |
| 7 | 0.50490 | 0.50730 | 2,261 | 28 | 2,233 | 1.24% | 1.04% | 11.21% | 95.46% | 66.57% | 28.88% |
| 8 | 0.50310 | 0.50490 | 2,260 | 16 | 2,244 | 0.71% | 0.60% | 11.27% | 96.05% | 77.84% | 18.22% |
| 9 | 0.50160 | 0.50310 | 2,261 | 5 | 2,256 | 0.22% | 0.19% | 11.33% | 96.24% | 89.16% | 7.08% |
| 10 | 0.50000 | 0.50160 | 2,260 | 101 | 2,159 | 4.47% | 3.76% | 10.84% | 100.00% | 100.00% | 0.00% |
| | | | 22605 | 2685 | 19920 | | | | | | 60.88% |

| | |
|---|---|
| Development GINI | 71.80% |
| Development KS | 60.88% |

**Table 7: Performance Results of Generalised Additive Models - II**

| Interval | Minmum Score | Maximum Score | Total Accounts | Number of goods | Number of bads | Good Rate | % goods | %bads | Cummulative goods : G(i) | Cummulative bads B(i) | KS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | 0 | 0 | | 0% | 0% | 0.00% | 0.00% | 0.00% |
| 1 | 0.6013 | 0.7311 | 2,260 | 1,177 | 1,083 | 47.90% | 5.88% | 41.59% | 5.88% | 41.59% | 35.71% |
| 2 | 0.5399 | 0.6013 | 2,261 | 1,592 | 669 | 29.59% | 7.96% | 25.69% | 13.84% | 67.28% | 53.44% |
| 3 | 0.5225 | 0.5399 | 2,261 | 1,936 | 325 | 14.37% | 9.68% | 12.48% | 23.52% | 79.76% | 56.24% |
| 4 | 0.5149 | 0.5225 | 2,260 | 2,093 | 167 | 7.39% | 10.46% | 6.41% | 33.99% | 86.18% | 52.19% |
| 5 | 0.5104 | 0.5149 | 2,261 | 2,159 | 102 | 4.51% | 10.79% | 3.92% | 44.78% | 90.09% | 45.31% |
| 6 | 0.5073 | 0.5104 | 2,261 | 2,196 | 65 | 2.87% | 10.98% | 2.50% | 55.76% | 92.59% | 36.83% |
| 7 | 0.5049 | 0.5073 | 2,260 | 2,218 | 42 | 1.86% | 11.09% | 1.61% | 66.85% | 94.20% | 27.35% |
| 8 | 0.5031 | 0.5049 | 2,261 | 2,240 | 21 | 0.93% | 11.20% | 0.81% | 78.05% | 95.01% | 16.96% |
| 9 | 0.5016 | 0.5031 | 2,261 | 2,248 | 13 | 0.57% | 11.24% | 0.50% | 89.29% | 95.51% | 6.22% |
| 10 | 0.5000 | 0.5016 | 2,260 | 2,143 | 117 | 5.17% | 10.71% | 4.49% | 100.00% | 100.00% | 0.00% |
| | | | 22606 | 20002 | 2604 | | | | | | 56.24% |

| | |
|---|---|
| **IN-time GINI** | 66.05% |
| **In-time KS** | 56.24% |

Tables 6 and 7 provide the performance results of generalised additive models, the model's rank order the risk monotonically across deciles. The model achieves a Max KS of 60.8/56.24 in development and in-time validation, the model achieves GINI of 71.80/66.05 in development and in-time validation.

**Table 8: Performance Results of Logistic Regression - I**

| Interval | Minmum Score | Maximum Score | Total Accounts | Number of goods | Number of bads | Good Rate | % goods | %bads | Cummulative goods : G(i) | Cummulative bads B(i) | KS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.60130 | 0.73110 | 2,260 | 1,277 | 983 | 56.50% | 47.56% | 4.93% | 47.56% | 4.93% | 42.63% |
| 2 | 0.53990 | 0.60130 | 2,261 | 658 | 1,603 | 29.10% | 24.51% | 8.05% | 72.07% | 12.98% | 59.09% |
| 3 | 0.52250 | 0.53990 | 2,260 | 349 | 1,911 | 15.44% | 13.00% | 9.59% | 85.07% | 22.58% | 62.49% |
| 4 | 0.51490 | 0.52250 | 2,261 | 163 | 2,098 | 7.21% | 6.07% | 10.53% | 91.14% | 33.11% | 58.03% |
| 5 | 0.51040 | 0.51490 | 2,260 | 95 | 2,165 | 4.20% | 3.54% | 10.87% | 94.67% | 43.98% | 50.70% |
| 6 | 0.50730 | 0.51040 | 2,261 | 72 | 2,189 | 3.18% | 2.68% | 10.99% | 97.36% | 54.96% | 42.39% |
| 7 | 0.50490 | 0.50730 | 2,261 | 37 | 2,224 | 1.64% | 1.38% | 11.16% | 98.73% | 66.13% | 32.60% |
| 8 | 0.50310 | 0.50490 | 2,260 | 17 | 2,243 | 0.75% | 0.63% | 11.26% | 99.37% | 77.39% | 21.98% |
| 9 | 0.50160 | 0.50310 | 2,261 | 11 | 2,250 | 0.49% | 0.41% | 11.30% | 99.78% | 88.68% | 11.09% |
| 10 | 0.50000 | 0.50160 | 2,260 | 6 | 2,254 | 0.27% | 0.22% | 11.32% | 100.00% | 100.00% | 0.00% |
| | | | 22605 | 2685 | 19920 | | | | | | 62.49% |

| | |
|---|---|
| **Development GINI** | **76.20%** |
| **Development KS** | **62.49%** |

**Table 9: Performance Results of Logistic Regression - II**

| Interval | Minmum Score | Maximum Score | Total Accounts | Number of goods | Number of bads | Bad Rate | % goods | %bads | Cummulative goods : G(i) | Cummulative bads B(i) | KS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.60130 | 0.73111 | 2260 | 994 | 1266 | 56.02% | 4.97% | 48.62% | 4.97% | 48.62% | 43.65% |
| 2 | 0.53990 | 0.60130 | 2261 | 1672 | 589 | 26.06% | 8.36% | 22.62% | 13.33% | 71.24% | 57.91% |
| 3 | 0.52250 | 0.53990 | 2261 | 1935 | 326 | 14.42% | 9.67% | 12.52% | 23.00% | 83.76% | 60.75% |
| 4 | 0.51490 | 0.52250 | 2260 | 2071 | 189 | 8.36% | 10.35% | 7.26% | 33.36% | 91.01% | 57.66% |
| 5 | 0.51040 | 0.51490 | 2261 | 2170 | 91 | 4.03% | 10.85% | 3.49% | 44.21% | 94.51% | 50.30% |
| 6 | 0.50730 | 0.51040 | 2261 | 2194 | 67 | 2.96% | 10.97% | 2.57% | 55.17% | 97.08% | 41.91% |
| 7 | 0.50490 | 0.50730 | 2260 | 2222 | 38 | 1.68% | 11.11% | 1.46% | 66.28% | 98.54% | 32.26% |
| 8 | 0.50810 | 0.50490 | 2261 | 2243 | 18 | 0.80% | 11.21% | 0.69% | 77.50% | 99.23% | 21.73% |
| 9 | 0.50150 | 0.50810 | 2261 | 2251 | 10 | 0.44% | 11.25% | 0.38% | 88.75% | 99.62% | 10.86% |
| 10 | 0.50000 | 0.50150 | 2260 | 2250 | 10 | 0.44% | 11.25% | 0.38% | 100.00% | 100.00% | 0.00% |
| | | | 22606 | 20002 | 2604 | | | | | | 60.75% |

| | |
|---|---|
| **Development GINI** | **75.41%** |
| **Development KS** | **60.75%** |

Tables 8 and 9 provide the performance results of logistic regression, the model's rank order the risk monotonically across deciles. The model achieves a Max KS of 62.4/60.75 in development and in-time validation, the model achieves GINI of 76.80/75.41 in development and in-time validation.

## Conclusion

In this paper, we have outlined generalised additive logistic models with their application for marketing response models. It is shown that generalised additive models perform equivalently well with logistic regression. As generalised additive model relaxes the assumption of linearity between the predictors and the response and avoids the problem of model misspecification, which is often prone to happen in generalised linear model, it is easily address by generalised additive models. Secondly, by incorporating non-linear effects, generalised additive model helps discover the hidden pattern of predictors and therefore improves performance on models when they are applied on larger datasets for scoring and also ensures that the models are stable over a period of time.

## References

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation Number. *Math.,* 31, 377-403.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models.* New York: Chapman & Hall.

Liu, W., & Cela, J. (2007). Improving credit scoring by generalized additive model. SAS global forum 2007 (paper 078-2007).

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. Chapman and Hall.

Moro, S., Laureano, R., & Cortez, P. (2011). *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology*. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

Muller, M. (2000). *Semi-parametric extensions to generalized linear models*. Habilitationsschrift.

Wood, S. N. (2006). Generalized additive model: An introduction with R. Chapman and Hall/CRC.

Stone, C. J. (1985). Additive regression and other non-parametric models. *Annals of Statistics,* 13, 689-705.

## Web Reference

http://support.sas.com/kb/32/927.html

Article can be accessed online at http://www.publishingindia.com

# Restaurant Recommendation System

**Surajit Ghosh Dastidar***

**Abstract**

In the present paper a restaurant recommendation system has been developed that a recommends a list of restaurants to the user based on his preference criteria. There are two kinds of data files that have been used: restaurant master and customer master. Restaurant master consists of restaurant specific data and customer master consists of customer specific data. We have used decision tree algorithm to classify the customers into high, medium and low budget buckets based on customer demographics and purchase behaviour variables. Similarly, restaurants are also classified based on price category. The rules given by the decision tree algorithm are fed into a dashboard designed using MS Excel. The user can use this dashboard to get a list of restaurants based on his individual preference. The restaurant list is sorted based on users location details with the closest restaurant coming at the top of the list.

**Keywords:** Restaurant, Recommendation System, Decision Tree

## Introduction

The concept of automated recommender systems was introduced in the 1990's and since then have revolutionised e-commerce by providing personalised recommendations and predictions over a variety of large and complex product offerings. A recommender system is any system that produces individualised recommendations as output or has the effect of guiding the user in a personalised way to interesting or useful objects in a large space of possible solutions (Burke, 2002). Recommendations made by such systems can help users navigate through large information about product descriptions, news articles or other items. Apart from user product preferences and taste, location based recommender system also exists which provides recommendation based on location preferences.

## Literature Review

Recommender systems are widely popular in e-commerce environment. For example, UrbanSpoon allows online users to see which restaurants have been visited frequently by friends and relatives. Users can also specify the type of cuisine and price range to select their preferred restaurants. Yelp provides location, user reviews to find their restaurant of choice. Rinner and Raubal (2004) designed Hotel Finder which recommends hotels based on users location, spatio-temporal constraints, and specification. Ringo music recommender system allows users to express their musical preferences by rating various artists and albums and also suggests based on similar preferences of other users. FourSquare recommends restaurants based on user-specified criteria such as rankings and ratings. Maes, Guttmann, and Moukas (1999) provides a recent survey of recommendation systems.

Recommender systems can broadly be categorised into the following:

### Collaborative or Social Filtering Recommender Systems

These systems aggregate data about customers purchasing habits or preferences, and make recommendations to other users whose profile matches with the past existing users. They basically use the concept of profiling. Collaborative systems works by building predictive models in order to predict the estimated of how much the user will like the set of items listed.

### Content based Recommender Systems

These systems recommend items based on user item description and user profile (demographic, purchase). Content-based recommendation systems are used in

* IMT Hyderabad, Hyderabad, Telangana, India. Email: sghoshdastidar@gmail.com

various domains ranging from recommending web pages, news articles, restaurants, television programs, and items for sale. It maintains the user profile in the database and updates the same on regular basis. These systems use supervised learning concept in order to predict the most likely results of a query given by the consumer. Content-based recommendation system analyse item description in order to identify items that are in line with the interest of the user.

### Knowledge based Recommender Systems

This kind of systems uses knowledge about the users and products to pursue a knowledge-based approach to generate recommendations for a particular user and product.

## Objective

The objective of this paper is to design and implement a localised, personalised, and content-based recommender system for restaurants. It should be easy to understand and implement.

## Data Source

The data was obtained from University of California Irvine machine learning repository. There were nine data files which were categorised into restaurants, customer and customer ratings.

### Data File for Restaurants

- Payment information
- Cuisine information
- Operating hours information
- Parking facility information
- Geographical and profile information

### Data Files for Customers

- Customer cuisine
- Customer payment
- Customer profile

### Data Files for Customer Ratings

Ratings given by the customers to each restaurant based on three parameters namely ratings, food and service.

Two master files namely restaurant master and customer master were created from the above mentioned eight files. Restaurant master file contained all the information regarding a particular restaurant. There were a total of 95 restaurants in the file. Customer master file contained all the information regarding a particular customer. There file had information of 135 customers. Table 1 gives the details of all the variables for each of the files.

**Table 1: Details of Master files**

| *Customer Master* | *Restaurant Master* |
|---|---|
| userID | placeID |
| Latitude | Rcusine |
| Longitude | Rpayment |
| Smoker | Parking |
| drink_level | Weekday time |
| dress_preference | Sat |
| Ambience | Sun |
| Transport | Latitude |
| marital_status | Longitude |
| Hijos | Name |
| birth_year | Alcohol |
| Interest | smoking_area |
| Personality | dress_code |
| Religion | Accessibility |
| Activity | Price |
| Color | Rambience |
| Weight | Franchise |
| Budget | Area |
| Height | other_services |
| Cuisine | |
| Payment | |

## Methodology

Literature review suggests that decision tree is the most preferred technique for content-based recommendation system. Hence, decision tree was chosen for categorising the customers and the restaurants. The decision tree algorithm was implemented using SAS Enterprise Miner 4.1

## Results

### Descriptive Analysis

It has been observed that 69% of the customers had a medium budget while only 4% had a high budget (Fig. 1). Mexican cuisine is offered by almost 29% all the restaurants (Fig. 2). 91% of the customers preferred cash transactions (Fig. 3). Around 45% of the restaurants did not have any parking facility (Fig. 4) while 33% offered alcohol services (Fig. 5). More than 60% of restaurants had smoking facilities (Fig. 6). The dress code for 90% of the restaurants was informal (Fig. 7).

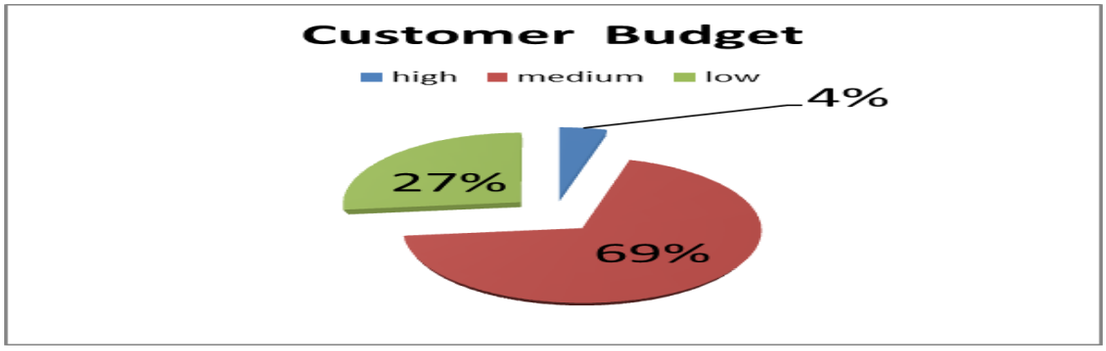



**Fig. 1: Customer Budget**

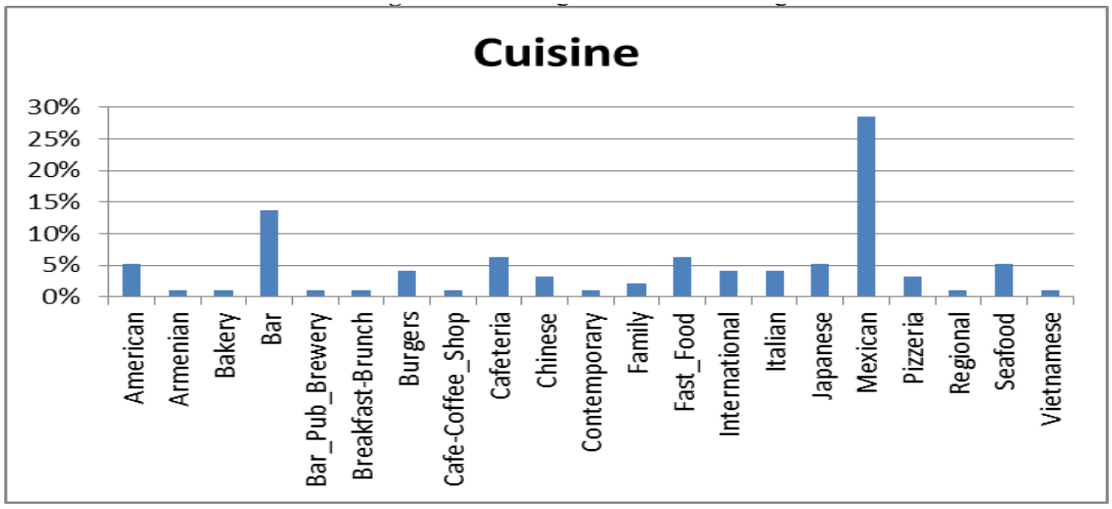



**Fig. 2: Cuisine**

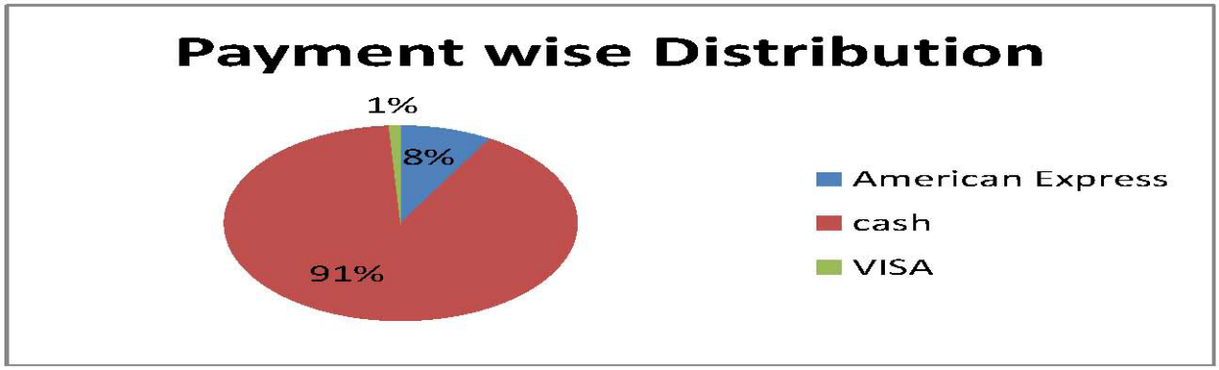



**Fig. 3: Payment wise Distribution**

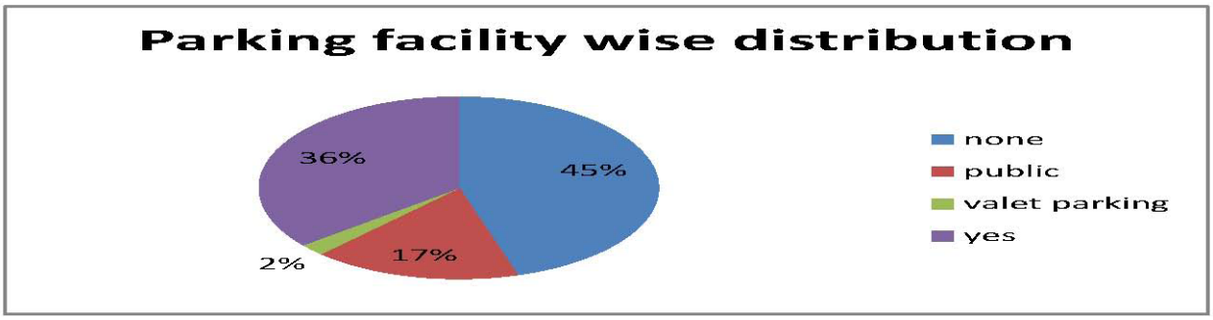



**Fig. 4: Parking Facility wise Distribution**

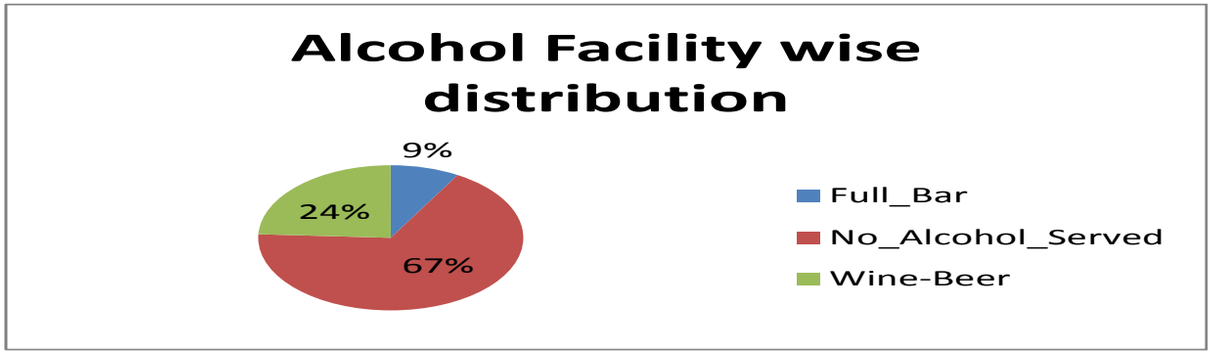



**Fig. 5: Alcohol wise Distribution**

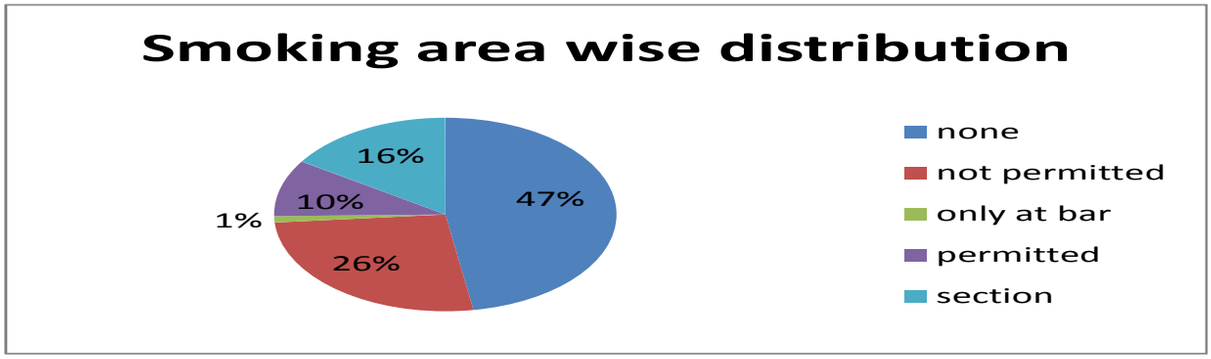



**Fig. 6: Smoking area wise Distribution**

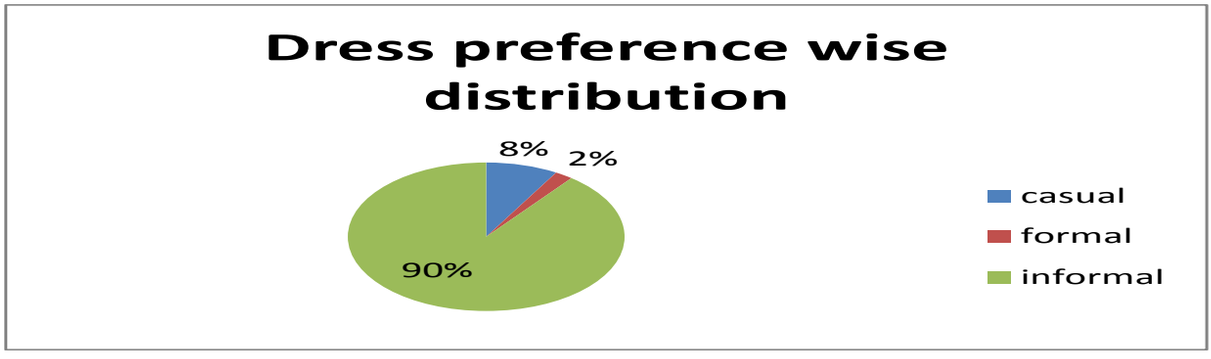



**Fig. 7: Dress Preference wise Distribution**

We also clustered the restaurants into three clusters based on the average user ratings for each of the restaurants. The following table summarises the results.

**Table 2: Ratings of Clusters of Hotels**

| Avg User Rating | Cluster No. | High | Low | Medium | Grand Total |
|---|---|---|---|---|---|
| 2.49 | 1 | 7% | 64% | 29% | 28 |
| 4.75 | 2 | 20% | 20% | 60% | 25 |
| 3.51 | 3 | 17% | 26% | 57% | 42 |

Cluster 1 had 64% restaurants in the low budget category. Interestingly, the average user ratings for this cluster was also found out to be 2.49. Cluster 2 had 80% of the restaurants in either high or medium priced category and had the maximum average user rating of 4.75. In cluster 3 the average user rating was 3.51 which constituted 74% of the restaurants belonging to medium or high priced restaurants. Hence, it can be concluded the majority of the low priced restaurants belonged to cluster 1 which also had a low user ratings while cluster 2 had the highest average user rating of 4.75. From this analysis we can conclude that low budget customer tend to go to low priced restaurants while high budget customer tend to go to high priced or medium priced restaurants. The first preference for medium budget customers is medium priced restaurants followed by low priced restaurants. In this study we have assumed a one to one relation between customer budget and restaurant category as stated earlier.

All the restaurants were classified into low, medium and high categories based on price. The restaurants were also classified into three categories "best", "ok", and "not ok" based on services offered, food quality, and review. It was observed that the number of "not ok" restaurants was maximum in the low price category compared to medium and high priced categories. It was interesting to note that none of the high priced category restaurants belonged to the "not ok" category.

The decision tree was obtained for the customers as shown in Fig. 8. The misclassification rate was found to be 0.20. Only those rules were considered for developing the dashboard where the probability of categorising the customers in any of the low, medium or high budget categories were greater than 70%

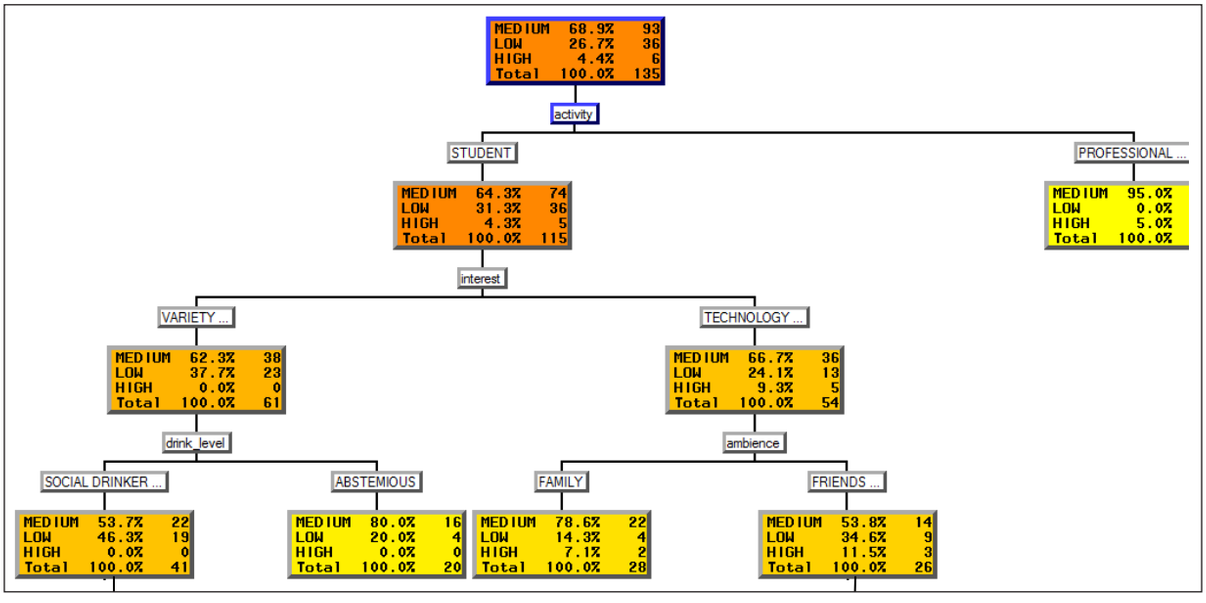



**Fig. 8: Decision Tree for Customers**

The decision tree was obtained for the restaurants as shown in Fig. 9. The misclassification rate of the following tree was found to be 0.32. We have considered only those rules where the probability of categorising the restaurants into low, medium or high priced restaurant is more than 0.70.

The decision rules will help the existing restaurant owners to identify what differentiates high priced restaurants from medium or low priced ones. Accordingly, the restaurant owner can add relevant features and functionalities to his existing restaurants in order to attract high budget customers.

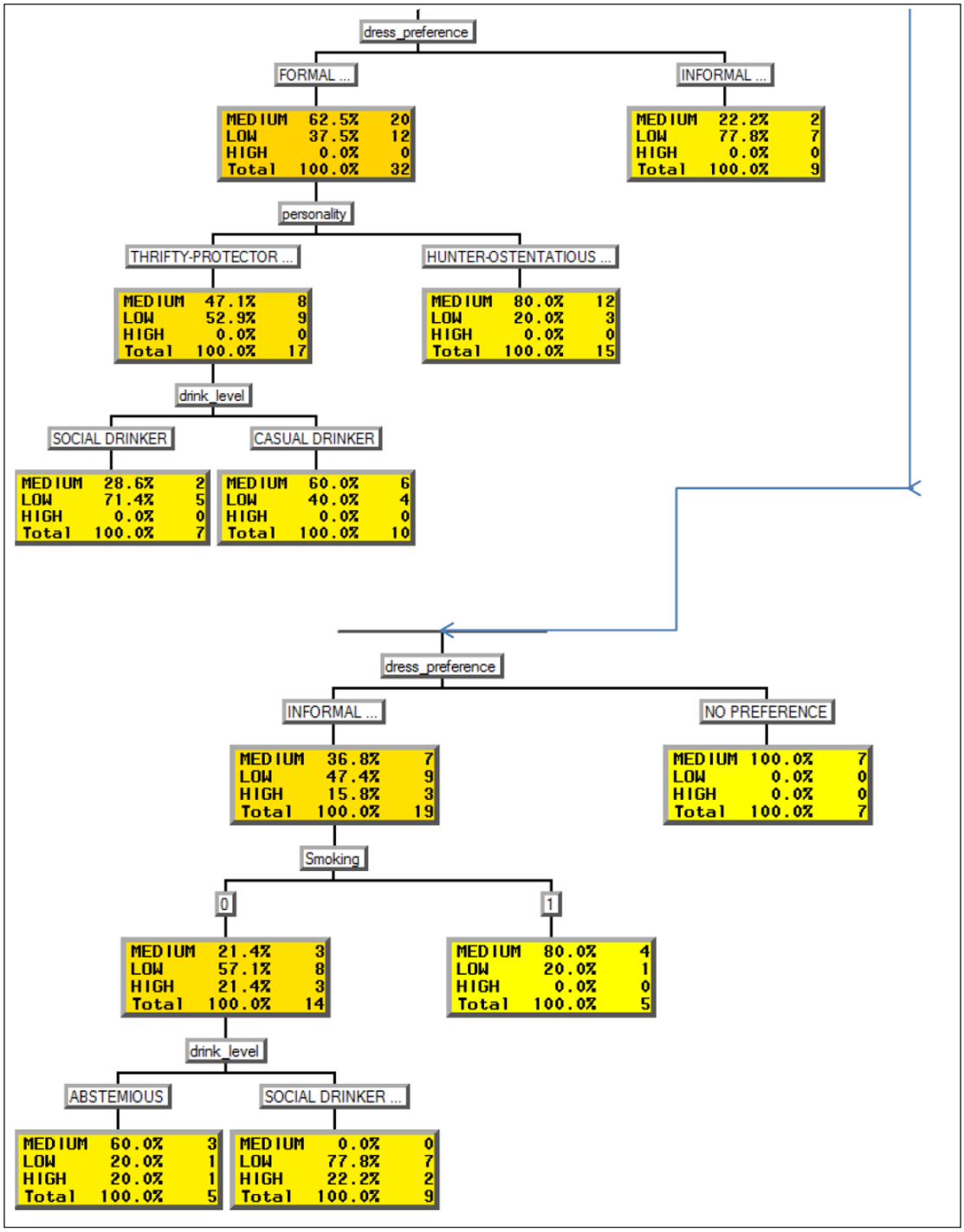



**Fig. 9: Decision Tree for Restaurants**

We have designed a dashboard in excel which is primarily meant for the customers. The customers can enter his personal preferences for a restaurant based on choices like smoking, interests, drink level, dress preference, ambience, personality and activity. The underlying decision tree rules will categorise the customer as high, medium or low budget customer and display a list of all the restaurants belonging to a particular category. If the customer further enters the location details of him in terms of latitude and longitude, the existing list of restaurants is further sorted in ascending order with the nearest restaurant coming at the top of the list.

Table 3 shows the input format for entering customer preferences. The customer preferences have to be entered in binary format, 1 meaning TRUE and 0 meaning FALSE.

**Table 3: Input Format for Entering Customer Preferences**

| Customer specifications input | | | | |
|---|---|---|---|---|
| Smoking | FALSE | TRUE | | |
| | 0 | 1 | | |
| | | | | |
| Interest | variety | retro | eco-friendly | technology |
| | 0 | 1 | 1 | 1 |
| | | | | |
| drink level | abstemious | casual drinker | social drinker | |
| | 0 | 0 | 1 | |
| | | | | |
| dress preference | formal | informal | no-preferecne | elegant |
| | 0 | 1 | 0 | 0 |
| | | | | |
| ambience | family | friends | solitary | |
| | 1 | 1 | 0 | |
| | | | | |
| personality | conformist | hard-worker | hunter-ostentatious | thrifty-protector |
| | 0 | 1 | 0 | 0 |
| Activity | student | proffesional | unemployed | workingclass |
| | 0 | 1 | | 0 |

| | *User Location details* |
|---|---|
| Latitude | 180 |
| Longitude | 102 |

Table 4 shows the output which is a list of restaurants along with restaurant details and distance from customer's existing location as specified by the latitude and longitude.

**Table 4: List of Restaurants along with Other Details**

| *Restaurant Recommendation system* | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Restaurant Recommendation system* | *Distance* | *Cuisine* | *Weekday* | *Saturday* | *Sunday* | *Parking* | *Area* |
| **Name** | *Distance* | *Cuisine* | *Weekday* | *Saturday* | *Sunday* | *Parking* | *Area* |
| Restaurant Bar Coty y Pablo | 1,941.04 | Bar | 16-24 | 16-24 | 16-24 | none | closed |
| El cotorreo | 1,943.37 | Family | 0-8 | 0-8 | 0-8 | none | open |
| Cafeteria cenidet | 1,944.12 | Cafeteria | 0-8 | 0-8 | 0-8 | public | closed |
| Log Yin | 1,946.28 | Mexican | 16-24 | 16-24 | 16-24 | yes | closed |
| Subway | 1,946.80 | Fast_Food | 16-24 | 16-24 | 16-24 | public | closed |
| McDonalds Centro | 1,946.90 | American | 16-24 | 16-24 | 16-24 | none | closed |
| TACOS CORRECAMINOS | 2,169.71 | Mexican | 16-24 | 16-24 | 16-24 | none | closed |
| TACOS EL GUERO | 2,170.56 | Mexican | 16-24 | 16-24 | 16-24 | none | closed |
| tacos de la estacion | 2,170.57 | Mexican | 0-8 | 0-8 | 0-8 | none | open |
| little pizza Emilio Portes Gil | 2,171.78 | Armenian | 16-24 | 16-24 | 16-24 | none | closed |
| palomo tec | 2,171.84 | Mexican | 16-24 | 16-24 | 16-24 | none | closed |
| tacos de barbacoa enfrente del Tec | 2,171.86 | Mexican | 8-16 | 8-16 | 8-16 | public | open |
| Carreton de Flautas y Migadas | 2,171.90 | Mexican | 0-8 | 0-8 | 0-8 | none | open |
| puesto de gorditas | 2,171.92 | Regional | 0-8 | 0-8 | 0-8 | public | open |
| tacos abi | 2,171.93 | Mexican | 16-24 | 16-24 | 16-24 | none | closed |
| Hamburguesas La perica | 2,172.16 | Mexican | 16-24 | 16-24 | 16-24 | public | open |

| | | *Distance(miles)* |
|---|---|---|
| Nearest Restaurant | Restaurant Las Mananitas | 1,947 |

## Conclusion

The paper discusses about the design and development of a localised, personalised and content-based recommendation system for restaurants which is easy to understand and implement using MS Excel. However, the location finder is limited in its functionalities since the location of the customer has to be entered manually. In an online or mobile environment the system can be designed to capture the location details from the internet address or the in-built GPS embedded in most of the mobile phones. In addition, the accuracy of the decision tree is limited to the small amount of customer data. Further testing needs to be done with larger database of customers and restaurants.

## References

Burke, R. (2002). Interactive critiquing for catalog navigation in e-commerce. *Artificial Intelligence Review*, 18(3-4), 245-267.

Maes, P., Guttmann, R. H., & Moukas, A. G. (1999). Agents that buy and sell, *Communications of the ACM*, 42(3), 81-91.

Rinner, C., & Raubal, M. (2004). Personalised multi-criteria decision based strategies based on location based decision support. *Journal of Location Geographic Information Sciences*, 10, 149-56.

## Web References

Foursquare.com

www.urbanspoon.com

Article can be accessed online at http://www.publishingindia.com

# Consumer Feedback Analysis Through Social Media for B2C Electronic Companies in India

**Riktesh Srivastava***

**Abstract**

Indian B2C electronic commerce market is rising at an aggressive pace of 21.3% and is likely to reach $28 billion revenue by 2019-2020 with annual growth rate of 45% in next 4 years. Also, the electronic commerce contributes 1.23% of the consolidated 7.6% GDP of India. The electronic commerce progression rate for India is expected to be 31.2%, as compared to 9.9% and 8.3% for China and Australia, respectively during 2016-2021. Also, B2C electronic commerce industry in India is the fastest growing industry, as matched to other industries, and has reached $38 billion market value in 2016, a jump of 67% from 2015. Also with mobile shopping further maturing and consumer mindshare continuing to split across multiple devices, these companies struggle to align consumer interactions with business strategies. It is due to this reason, they use social media for better consumer interactions and spreading brand awareness digitally. It is presumed that social media has the ability to increase sales because of their strong online presence. Also, when these companies communicate with consumers through social media networks, they are able to get feedback instantly, which gives them quick acumen into what they want. The current study focuses on an analysis of these feedbacks collected by top 5 B2C electronic companies in India, namely, Amazon India, Flipkart, Snapdeal, Myntra, and eBay India. The feedback analysis is conducted based on the tweets from these companies on Twitter for 3 months, from 01-01-2017 to 31-03-2017. The experiment is conducted using Naïve Bayes Algorithm for 1500 tweets and places the response into one of the quadrants on proposed investigation model called "4AIM" – 4A Investigation Model. Based on the outcomes, the study adopts the generic social media strategies (BWDC, 2014), which these companies can embrace and implement accordingly.

**Keywords:** Electronic Commerce, Naïve Bayes Algorithm, 4AIM, Amazon India, Flipkart, Snapdeal, Myntra, eBay India, Social Media Strategies

## Introduction

For B2C electronic commerce companies, it is difficult to identify and influence the factors that drive consumers' attitudes and behaviour. Conventionally, in order to get consumer insights and feedbacks, these companies trusted on a blend of quantitative data from surveys (to evaluate consumer satisfaction and feedback) and qualitative insights from focus groups and interviews. However, both types of tools relied deeply on consumers' remembrances and recall capability, which declines hastily. It was due to this motive, Internet-based research tools were introduced to capture consumer experiences almost instantly. However, these tools provided just 15% of consumers' encounters with companies (Emma & Macdonald, 2012). Advent of social media has both motivated and accorded a dramatic change the way businesses and consumers interact. Social sites such as Twitter and Facebook provides platform as an integrated communication model, where consumers have the choice of how and when they communicate with companies (Causon, 2015). Nielsen reported that nearly 70% of adults who use social media to buy products digitally (Neilsen, 2012). Another study states that 44% businesses had acquired consumers using Twitter (Georgieva, 2012). Thus, the most important usage of Twitter by Electronic Commerce companies are consumer interaction (Blacknell, 2011) and audience extension (Booth & Matic, 2011).

The study is divided into four steps mentioned in Fig. 1:

* Associate Professor, Information Systems, Skyline University College, Sharjah, UAE. Email: rsrivastava@skylineuniversity.ac.ae

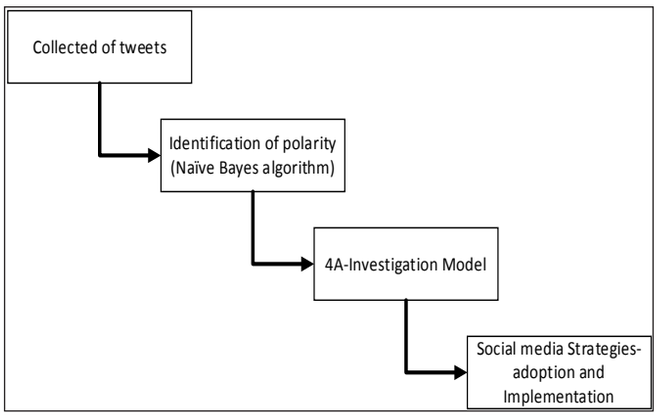



**Fig. 1: Step-by-step Illustration of the Study**

As mentioned in Fig. 1, step 1 collected the tweets to maximum of 1500 and is mentioned in second section. Third section elaborates step 2 and exemplifies use of Naïve Bayes algorithm for feedback analysis. Step 3 mentions recommended 4A-Investigation Model (4AIM) and places the positive polarity of feedback collected into one of quadrants, as stated in fourth section. Step 4 outlines the social media strategies to be adopted and implemented by these companies and is mentioned in fifth section.

## Collection of Tweets

During the study, it was witnessed that these companies had two Twitter accounts (except for eBay). One is the official account, where these companies displayed the updates, sale, and offers, and, other for support or assist consumers for the queries. Table 1 mentions the comprehensive status of twitter accounts of these companies.

**Table 1: Twitter Status (as on 31/03/2017)**

| *Company* | *Twitter Account(s)* | *Total Tweets* | *Total Followers* |
|---|---|---|---|
| Amazon | @amazonIN | 21.7K | 637K |
| | @AmazonHelp | 1.31M | 103K |
| Flipkart | @Flipkart | 32.8K | 1.48M |
| | @Flipkartsupport | 332K | 63.4K |
| Snapdeal | @Snapdeal | 26.2K | 696K |
| | @Snapdeal_Help | 217K | 24K |
| Myntra | @Myntra | 80.9K | 350K |
| | @MyntraSupport | 29.4K | 16.7K |
| eBay | @ebayindia | 84K | 210K |

Step 1 executes the R code to collect the tweets and comments from 01-01-2017 to 31-03-2017 to maximum of latest 1500 tweets. The outcome of step 1 is given in Table 2.

**Table 2: Feedback Collected**

| *Twitter Account(s)* | *Tweets (n=1500)* |
|---|---|
| | *Feedback Collected* |
| @amazonIN | 1500 |
| @AmazonHelp | 1500 |
| @Flipkart | 1500 |
| @Flipkartsupport | 1500 |
| @Snapdeal | 1500 |
| @Snapdeal_Help | 1500 |
| @Myntra | 1500 |
| @MyntraSupport | 818 |
| @ebayindia | 1199 |

## Identification of Polarities Using Naïve Bayes Algorithm

Naïve Bayes algorithm is used to outline the contextual polarity of comments by consumers of electronic commerce companies. The comments are collected as "bag of words" and provided to Naïve Bayes algorithm, which treats each comment independent of each other. Based on each word from each tweet, the algorithm determines the classes of each word as positive, neutral, or negative. The aggregate of class for each tweet then classifies into one of three polarities.

The mathematical representation of Naïve Bayes algorithm is represented in equation 1 as:

$$P(A/B) = \frac{P(B|A)P(A)}{P(B)} \quad \ldots(1)$$

where,

$P(A|B)$ is the probability of A (class), given B (tweet).

$P(B|A)$ is the probability of B (tweet), given A (class).

$P(A)$ is the probability of A (class), and is independent of each other.

$P(B)$ is the probability of B (tweet), and is independent of each other.

Based on equation (1), positive and negative tweet are represented as

$$P_{(\text{positive}|\text{tweet})} = \frac{P(\text{tweet}|\text{positive})P(\text{positive})}{P(\text{tweet})} \quad \ldots (2)$$

$$P_{(\text{positive}|\text{tweet})} = \frac{P(\text{tweet}|\text{negative})P(\text{negative})}{P(\text{tweet})} \quad \ldots (3)$$

It is observed that probability of tweets, $P(\text{tweet})$ is constant, and can thus be ignored. Thus, equations (2) and (3) can be represented as:

$P(\text{positive tweet}) =$

$P(\text{tweet}|\text{positive})\ P(\text{positive})$ … (4)

$P(\text{negative}|\text{tweet}) =$

$P(\text{tweet}|\text{negative})\ P(\text{negative})$ … (5)

The more precise notation of each class is thus given in equations (6), (7), and (8) respectively.

$$P(\text{positive}) = \sum_{j=1}^{m}\sum_{i=1}^{n} P(T_i|\text{positive}) \quad \ldots (6)$$

$$P(\text{positive}) = \sum_{j=1}^{m}\sum_{i=1}^{n} P(T_i|\text{negative}) \quad \ldots (7)$$

$$P(\text{neutral}) = 1 - [P(\text{positive}) + P(\text{negative})] \quad \ldots (8)$$

where,

$i = 1,.\ n \rightarrow$ à total number of words for each tweet

$j = 1,.\ m \rightarrow$ à total number of tweets

Based on equations (6), (7), and (8), Table 3 gives the polarities of tweets for these companies.

**Table 3: Polarity Status of Selected Companies**

| *Twitter Account(s)* | *Polarity* | | |
|---|---|---|---|
| | + | +/- | - |
| @amazonIN | 66.87% | 14.80% | 18.33% |
| @AmazonHelp | 54.87% | 17.73% | 27.40% |
| @Flipkart | 66.60% | 8.20% | 25.20% |
| @Flipkartsupport | 46.67% | 20.00% | 33.33% |
| @Snapdeal | 56.20% | 16.73% | 27.07% |
| @Snapdeal_Help | 54.93% | 18.87% | 26.20% |
| @Myntra | 76.67% | 11.67% | 11.67% |
| @MyntraSupport | 72.13% | 14.67% | 13.20% |
| @ebayindia | 78.32% | 12.93% | 9.17% |

The twitter graphs are constructed for the companies in stages using publicly available data from the Twitter API. From the list of each companies' tweets, only the comments on which the consumers react are collected; this cuts unknown consumers' who did not comments and thus are unlikely to provide useful information. Also, the rapid pace of growth on Twitter, the polarity tends to grow quickly; thus the overall polarity is a representation of the companies' current social status and not the exact status that existed at the time of the tweet. The feedbacks from these consumers are collected for the period 01-01-2017 to 31-03-2017 and maximum 1500 tweets were collected.

Figs. 2 to 6 show the feedback polarity breakdown for these companies.

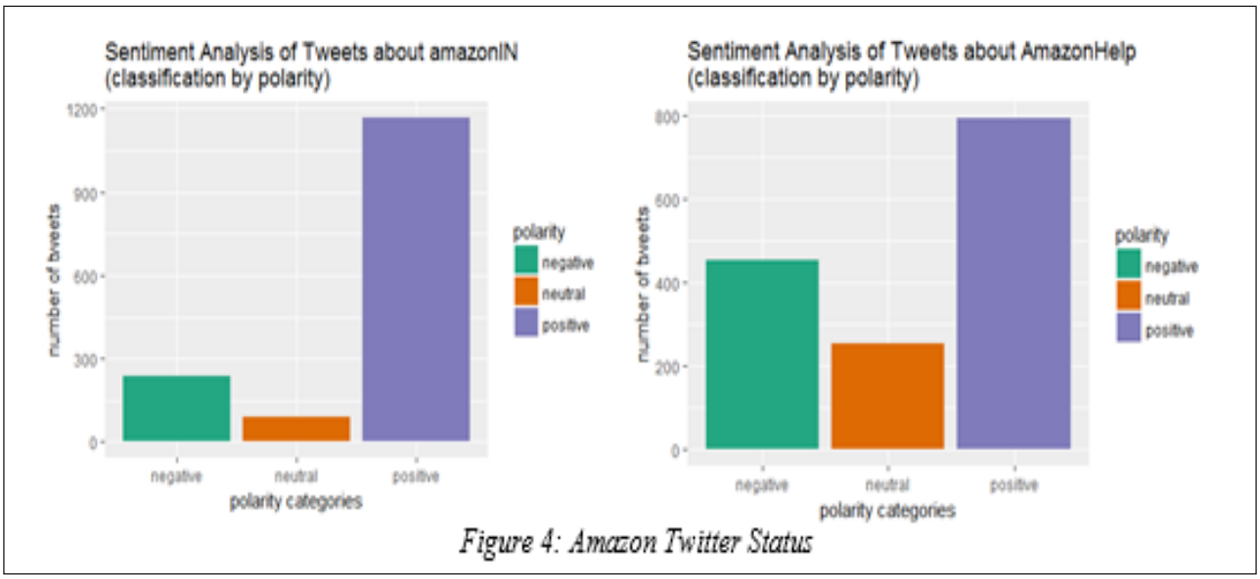



*Figure 4: Amazon Twitter Status*

**Fig. 2: Amazon Twitter Status**

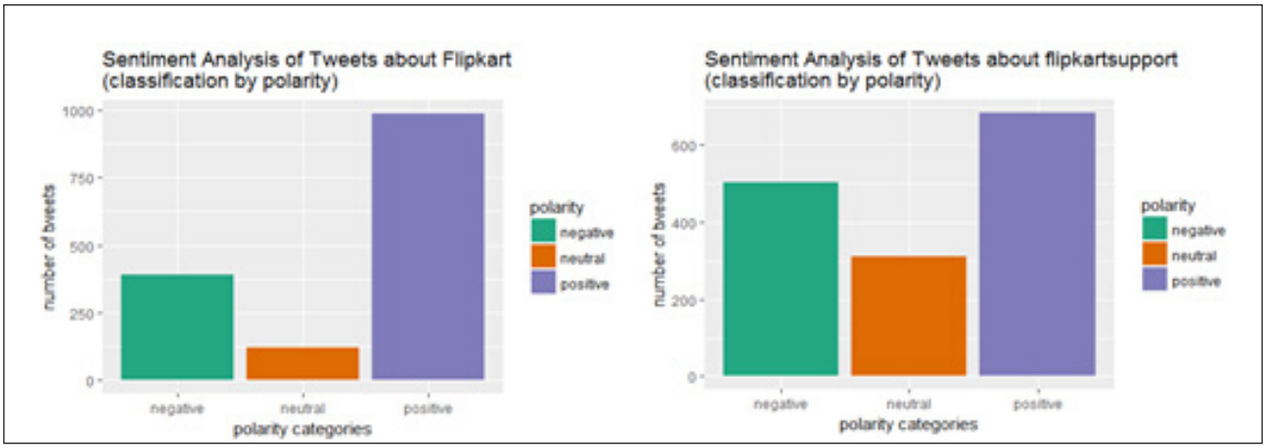



**Fig. 3: Flipkart Twitter Status**

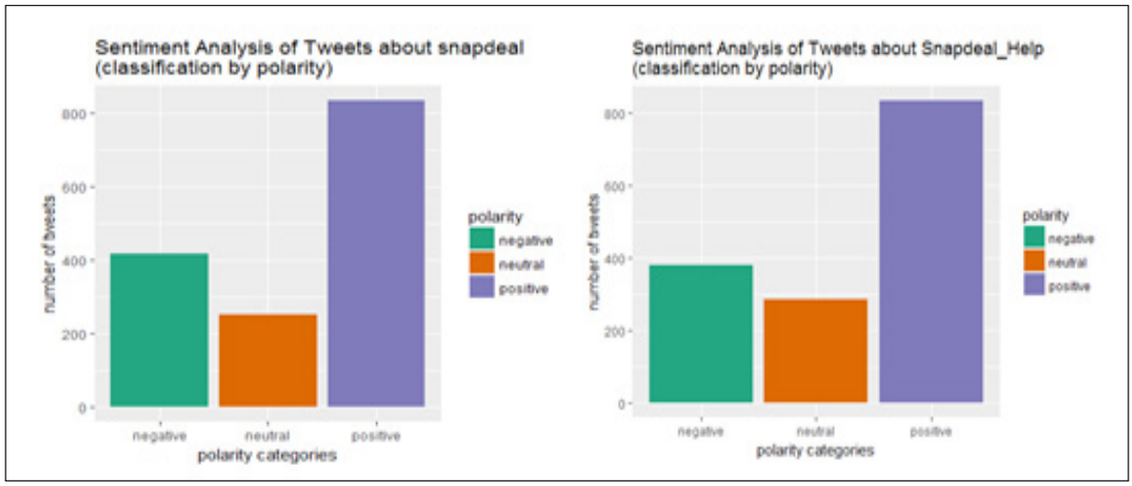



**Fig. 4: Snapdeal Twitter Status**

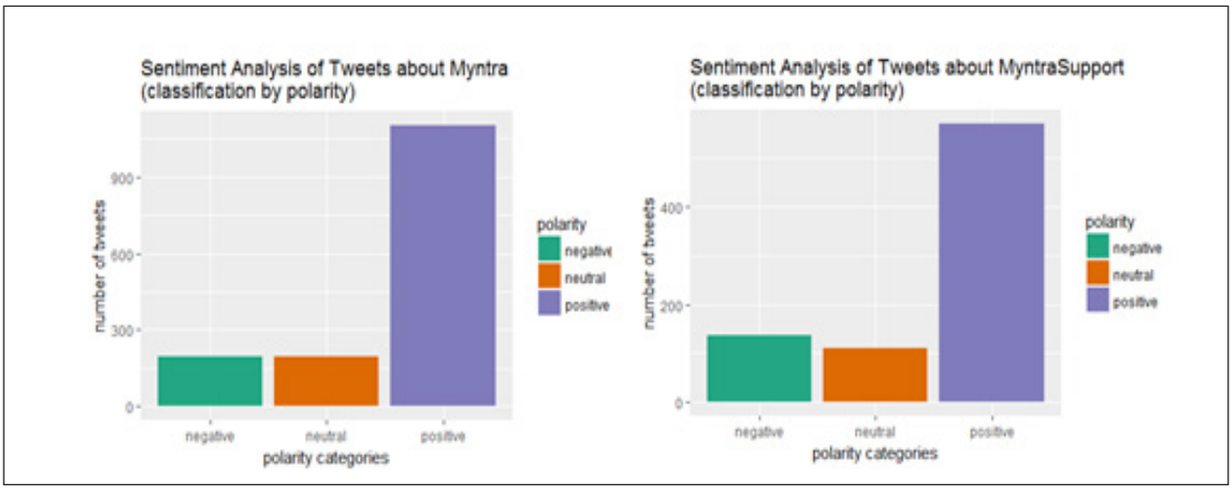



**Fig. 5: Myntra Twitter Status**

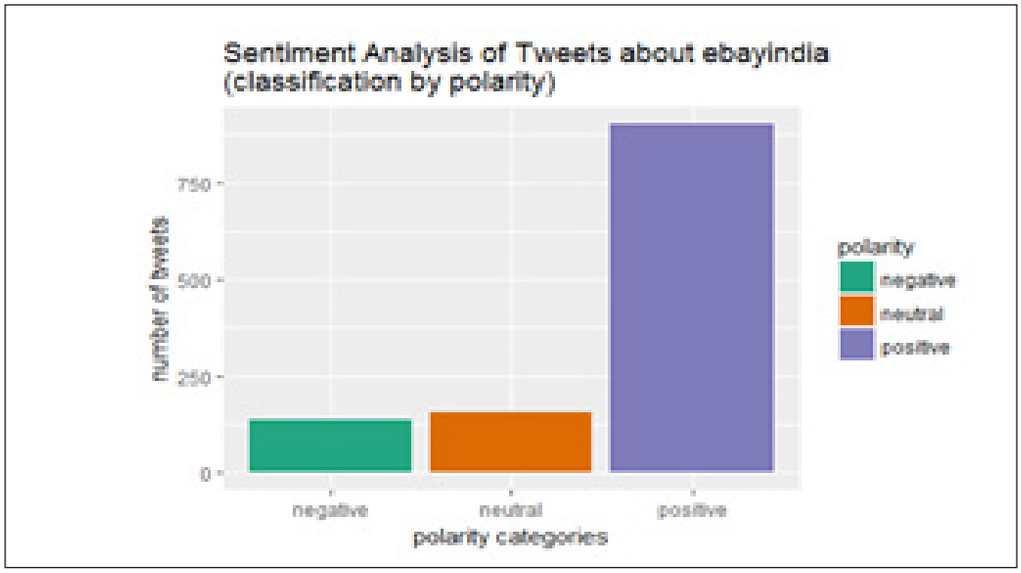



**Fig. 6: eBay India Twitter Status**

## 4A-Investigation Model

4A – Investigation Model (4AIM) is proposed to analyse the consumer feedback. The model uses positive polarity to identify the status of current engagement of these companies with consumers, as illustrated in Table 4.

**Table 4: Evaluation Table for 4A-Investigation Model**

| *Positive Polarity (in %)* | | *4AIM states* |
|---|---|---|
| 0 | 30 | Anxious |
| 31 | 60 | Apart |
| 61 | 80 | Ardent |
| 81 | 100 | Active |

The analysis for a model is mentioned in Fig. 7. The model is divided into four quadrants based on the percentage of positive polarities. Placing the outcomes into these quadrants easily identifies the current state of social media adoption, and strategies to be adopted in case required.

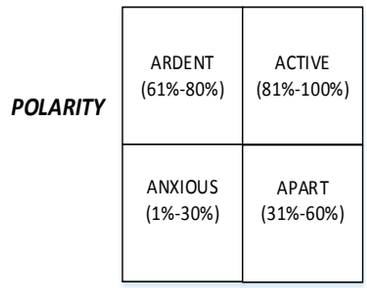



**Fig. 7: Feedback Analysis**

## Implication of 4AIM for the Observed Outcomes

This section illustrates the implication of outcomes in the proposed 4AIM. Fig. 8 displays the outcomes of the experiment conducted.

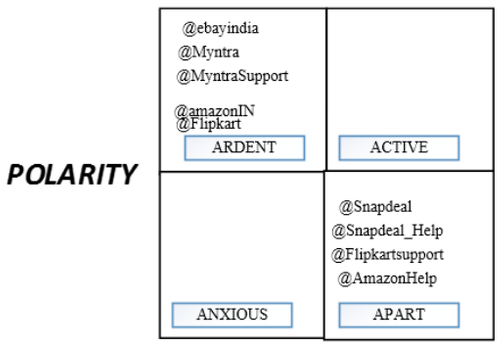



**Fig. 8: Outcomes of Twitter Status**

Surprisingly, the feedback for tweets are only in two states – apart and ardent. Also, none of these companies fall in "anxious" state, which demonstrates these companies have accepted Twitter and use it for updates and feedback quite regularly. However, surprisingly, none of these companies have reached "active" state, even after years of Twitter adoption, which is shocking.

Also, the average feedback for "ardent" state is 72.12% and for "apart" state is 53.17%. The study also specifies

that consumers are not contented with the response they are getting online from these companies. Thus, the help/support segment of Twitter account by these companies are not aiding consumers. Social media strategies should be in place to guide these companies to wrestle the consumer queries and respond accordingly. Fifth section exemplifies these strategies in element.

## Social Media Strategies for Electronic Commerce Companies

Social media strategies outline the detailed plan for these companies to contact potential consumer and device communication blueprint. For electronic commerce companies, increased communication through social media that will guide them to be in the "active" state of 4AIM.

Table 5 delineates the social media strategies to be adopted and instigated by these companies.

**Table 5: Social Media Strategies**

| *Strategy* | *Description* |
|---|---|
| S#1 (Know where your consumers are?) | Identify which platform consumers are engaged in. |
| S#2 (Engage consumers) | Open dialogue with consumer, content development, and consumer stories |
| S#3 (Build trust) | Genuine willingness to help |
| S#4 (Add social sharing buttons) | Social sharing buttons can be included on the website, so consumers can share them with others. |
| S#5 (Create Videos) | Videos of consumers with products, office environment to be updated in Social media. |
| S#6 (Exclusive offers) | The inclusion of Exclusion offers like free delivery, Upcoming sale, breaking news to be included. |
| S$7 (Competitors Analysis) | Three steps include:<br>Type of content<br>Social networks used along with number of followers and their interaction<br>Promotion strategies by competitors |
| S#8 (Don't always push products and promotions) | Blog on electronic commerce site and feed the blog content into social accounts<br>Share stories and messages from other sources<br>Pictures and videos of company events or engagements<br>Ask questions, discussion forums and poll using social media |
| S#9 (Infographics Investigation) | Generates high-value backlinks and helps in SEO |
| S#10 (Serve consumers) | Extend to social media to know more about consumers' satisfaction, problems and complaints. |

## Alignment of Social Media Strategies with 4AIM Model

After the identification of social media state through 4AIM, the use and enactment of strategies (as mentioned in Table 5) becomes straightforwardly fathomable.

Table 6 demonstrates the strategies to be agreed for different states of 4AIM for these companies.

**Table 6: Social Media Strategies for Different States of 4AIM**

| *4AIM State* | *Strategies to be adopted* |
|---|---|
| Anxious | S#1 (Know where your consumers are?) S#2 (Engage consumers) S#3 (Build trust) |
| Apart | S#4 (Add social sharing buttons) S#5 (Create Videos) S#6 (Exclusive offers) |
| Ardent | S$7 (Competitors Analysis) S#8 (Don't always push products and promotions) |
| Active | S#9 (Infographics Investigation) S#10 (Serve consumers) |

Social media offers organisations with a way to connect with their consumers. Customer service is a basic aspect and an obvious customer loyalty opportunity. 65% of users are willing to make more purchases from a brand if they get customer service on social networks (Carter, 2016). The recommendations for these companies are listed in Table 7.

**Table 7: Recommendations for Companies**

| *Twitter* | *Strategies' in place* |
|---|---|
| @amazonIN | S#7,S#8 |
| @AmazonHelp | S#4,S#5,S#6 |
| @Flipkart | S#7,S#8 |
| @Flipkartsupport | S#4,S#5,S#6 |
| @Snapdeal | S#4,S#5,S#6 |
| @Snapdeal_Help | S#4,S#5,S#6 |
| @Myntra | S#7,S#8 |
| @MyntraSupport | S#7,S#8 |
| @ebayindia | S#7,S#8 |

## References


Assocham. (Jan. 1, 2016). *E-Commerce Industry will cross $38 bln mark by 2016; Indian e-commerce market set to grow by 67% in 2016: Study*. Retrieved from http://www.assocham.org/newsdetail.php?id=5427

Blacknell, A. (2011). Using social media to drive engagement. *Strategic Communication Management,* 15(4), 13.

Booth, N., & Matic, J. A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions. *Corporate Communication,* 16(3), 184-191.

BWDC. (Feb. 22, 2014). *Useful Social Media Marketing Strategies for E-Commerce Websites*. Retrieved from http://blog.bangalorewebdesigncompany.com/useful-social-media-marketing-strategies-e-commerce-websites/

Carter, B. (Aug. 18, 2016). *Customer Loyalty Statistics: 2016 edition*. Retrieved from http://blog.accessdevelopment.com/customer-loyalty-statistics-2016-edition#loyalty

Causon, J. (2015). *Customer complaints made via social media on the rise.* The Guardian.

Emma, K., & Macdonald, H. N. (Sep. 2012). Better customer insight-in real time. *Harvard Business Review*.

Foundation, E. (2016). *India B2C E-commerce Report 2016.* Amsterdam: ECommerce Foundation.

Georgieva, M. (2012). *20 Revealing Stats, Charts, and Graphs Every Marketer Should Know*. Retrieved from https://blog.hubspot.com/blog/tabid/6307/bid/32985/20-Revealing-Stats-Charts-and-Graphs-Every-Marketer-Should-Know.aspx#sm.000012xpxjgf0rd2jxjmbj2fxjvfh

Lachhiramka, B. (Sep. 17, 2015). *Why is it important for ecommerce companies to invest in customer interaction management systems?* Retrieved from https://inc42.com/buzz/customer-interaction-management/

Neilsen. (2012). *Consumer trust in online, social and mobile advertising grows.* Neilsen.

PTI. (Sep. 1, 2016). *Indian e-commerce market could reach $28 billion by FY 2020.* Retrieved from The Economic Times: http://economictimes.indiatimes.com/industry/services/retail/indian-e-commerce-market-could-reach-28-billion-by-fy-2020-kotak-institutional-equities/articleshow/54249855.cms

PwC. (2015). *eCommerce in India:Accelerating growth.* PwC.

Satish Meena, V. S. (2017). *Forrester Data: Online Retail Forecast, 2016 To 2021 (Asia Pacific).* Forrester.

Top 10 ECommerce companies in India. (March 14, 2017). India. Retrieved from Companies in India: http://companiesinindia.net/top-10-ecommerce-companies-in-india.html.

Article can be accessed online at http://www.publishingindia.com

# Replenishment Policy in a Two-Echelon Supply Chain: An Analysis Using Discrete-Event Simulation

**Ruchir Prasoon[*], Maulik Agarwal[**], Ajith Kumar J.[***]**

**Abstract**

The present study identifies optimal inventory policies for two-echelon systems under the effects of supply disruptions and stochastic demand. Previous research has incorporated and addressed stochastic demand quite extensively but the study of supply disruptions is relatively new. Presently, supply disruptions are studied heuristically and through theoretical analysis to identify optimum inventory levels. The current study uses discrete event simulation to arrive at optimal policies under varied levels of supply disruptions and stochastic demand. It uses a designed experiment to vary disruption length and disruption frequency. We find that under conditions of supply disruption, a decentralised policy is more likely to yield lower costs than a centralised policy when disruption levels are high but a centralised policy is better otherwise.

**Keywords:** Supply Disruptions, Discrete-Event Simulation, Two-Echelon System, Experimental Design

## Introduction

Much research on supply chain and inventory models has focused on demand uncertainty. However, there has been a growing interest in supply uncertainty amongst academics and practitioners, particularly over the last decade (Snyder, Atan, Peng, Rong, Schmitt, & Sinsoysal, 2016). Supply uncertainty occurs when either a company or its suppliers, or both, are not able to deliver the required quantity at the right time (Al-Rifai & Rossetti, 2007. It is broadly classified into three categories, the first of which is supply disruption, that is, when the organisation or the suppliers cannot deliver or operate until the disruption is over. The second is yield uncertainty, where the quantity supplied by the supplier or organisation is subject to variation and uncertainty, while the third is stochastic lead-times wherein the amount of time to supply the goods is subject to variation (Arreola-Risa & DeCroix, 1998).

In this study, we considered the first kind, that is, supply disruptions. Disruptions in a supply chain can occur due to a variety of causes such as natural disasters, political or social unrests, strikes, and accidents. Even small disruptions in a supply chain can have a debilitating impact on the production and supply of goods and services of more than one firm. Thus, though a given type of disruption may be rare in the sense that it may occur less than once a year, it is important to be aware of the harm it can cause when it does occur, and to take appropriate preventive measures. Though some disruptions might be preventable, some are not under control and hence effective planning to mitigate or minimise the damage becomes crucial for an organisation. The effects of disruptions may last a very long time and even minor disruptions can have significant effects.

We modelled a two-echelon system, or more specifically, a one-warehouse multiple-retailer system (OWMR). An earlier work on OWMR systems is that of Atan and Snyder (2012). Two-echelon system can experience

[*] PGDBM Student, XLRI Xavier School of Management, Jamshedpur, Jharkhand, India. Email: b15043@astra.xlri.ac.in
[**] PGDBM Student, XLRI Xavier School of Management, Jamshedpur, Jharkhand, India. Email: b15149@astra.xlri.ac.in
[***] Professor, XLRI Xavier School of Management, Jamshedpur, Jharkhand, India. Email: akm@xlri.ac.in

more severe effects because of disruptions than single echelon systems. We explore whether a centralised or de-centralised system will be better for an organisation in the longer run by comparing costs and cost-variances for centralised and de-centralised system with optimal costs and variance under given simulation conditions for both scenarios, that is, with and without supply disruptions. In doing all this, we developed a discrete event simulation model and used it to conduct a series of experiments. Essentially, we quantified disruption risk using discrete event simulation and a few simplifying assumptions. The results from our analyses can help identify effective strategies under disruptions and the parts of supply chains to be focused on for minimising their effects.

The rest of this paper is as follows. In the following section, we present a review of the literature on supply chain disruptions and present the specific model addressed by this study. Then we discuss the simulation models built and the experiments conducted with the same. Subsequently, we present the results of the analyses and implications drawn from the same. The paper concludes by highlighting the key findings and suggests some directions for the future.

## Literature Background

A detailed literature review of OR/MS models of supply chain disruptions is offered by Snyder *et al*. (2016). This work reviews supply disruptions against the broad backdrop of supply uncertainty and discusses common modeling approaches. Some early authors in this area focused on risk in a JIT system, highlighting the importance of sharing the risk throughout the supply chain (e.g. Simchi-Levi, Snyder, & Watson, 2002) and keeping reserves of inventory to protect against disruptions (e.g. Sheffi, 2001).

Tang (2006) identifies two types of supply chain risk strategies: those that increase a supply chain's efficiency and those that increase its resilience. Efficiency is about a firm's operational ability to handle a disruption, while resilience is the ability of a firm to sustain operation and recover quickly, in the face of a disruption. Sheffi (2005) focuses on the latter and suggests that using local suppliers may be safer due to decreased transportation times and lengths, even though they may not quote the lowest unit costs. According to Tang and Tomlin (2008) a small amount of flexibility in a supply chain can have large payoffs if disruptions occur.

There have been both analytical as well as simulation-based approaches to model supply chain disruptions. Atan and Snyder (2012) study a two-echelon, one-warehouse, multiple retailer (OWMR) distribution system subject to supply disruptions. They build analytical models of the supply chain system and propose algorithms to find the optimal stocking levels of all locations in the system. They assume periodic review base-stock policies and deterministic demands at the retailers. Qi, Shen, and Snyder (2010) model an integrated supply chain design problem that determines the locations of retailers and the assignments of customers to retailers. Their aim is to minimise the expected costs of location, transportation, and inventory. The system is subject to random supply disruptions that may occur at either the supplier or the retailers and these authors demonstrate numerically that the cost savings from considering supply disruptions at the supply chain design phase, rather than at the tactical or operational phase, are usually significant.

Rong, Atan, and Snyder (2015) study continuous-review distribution systems with Poisson customer demands under a first-come, first-served allocation policy and develop heuristics to approximate the base-stock levels of all the locations in the system and discuss the strengths and limitations of these heuristics. Schmitt, Sun, Snyder, and Shen (2015) study a multi-location supply chain system in which supply is subject to disruptions and examine expected costs and cost variances in centralised and decentralised systems. They demonstrate that when demand is deterministic and supply may be disrupted, using a decentralised inventory design reduces cost variance through the risk diversification effect, and therefore a decentralised inventory system is optimal. When demand is stochastic and supply may be disrupted, they suggest that a risk-averse firm should typically choose a decentralised inventory system design.Some two-echelon models allow for disruptions only at the most upstream locations, that is the suppliers (e.g. Bollapragada *et al*., 2004, Boute *et al*., 2009), while some allow for disruptions at both echelons of a one-warehouse multiple-retailer (OWMR) system (e.g. Bulut and Snyder, 2009, Atan and Snyder, 2012).

Snyder *et al*. (2016) assess the effect of supply disruptions on inventory decisions by considering deterministic demand in a two-echelon system. Through theoretical analysis and numerical study, they obtain and propose the solution to find optimum stocking level under different

disruption conditions. They also assess the effect of disruptions if they occur close to the customer. This work considers three cases for disruptions, disruptions occurring at warehouse only, disruptions occurring at retailers only and disruptions occurring at both warehouse and retailers. Schmitt *et al*. (2014) study the cost and cost variances for the two inventory policies of centralisation and decentralisation under supply disruptions, in particular, the classical effects of risk pooling and risk diversification using an analytical approach and numerical study. Under some simplifying assumptions, the paper proposes decentralisation as the optimal policy under supply disruptions as the risk diversification effect prevails but notes that under stochastic demand and deterministic supply centralisation may be a better policy due to risk pooling effect.

For models considering more than two echelons, Hopp and Liu (2006) model an assembly system where disruptions may occur at any location in the network. Schmitt (2011) also considers a multi-echelon system where any stage may be disrupted, focusing on a combined serial-distribution system.

In contrast to these analytical approaches, a study that uses simulation is Deleris and Erhun (2005) who build a Monte Carlo model, while Snyder and Shen (2006) use discrete-event simulation to contrast supply chain uncertainty and demand uncertainty in optimal system design. Schmitt and Singh (2009) use a combination of Monte Carlo and discrete-event simulation to model downtime due to disruptions. In a later study, Schmitt and Singh (2012) use a discrete-event simulation based approach to demonstrate how system resilience can be improved by focusing on a supply chain network as a whole. They analyse inventory placement and back-up methodologies in a three echelon network and view their effect on reducing supply chain risk. They focus on risk from both supply disruptions and demand uncertainty and compare their impacts and mitigating strategies. A simulation model developed to capture an actual network for a consumer packaged goods company is used for the analysis. They present analysis and insights for multi-echelon networks and show how network utilisation and proactive planning enable reductions in supply chain disruption impact.

The current study considers an OWMR system with both supply-side disruptions as well as stochastic demand. Further, we also consider lead time to be uncertain or stochastic in nature. As analytical modeling with these conditions can be quite complex, the current study uses discrete-event simulation.

Summarising the discussion so far, the specific questions that this study aims to address are:

1. How does centralisation compare with decentralisation on cost of backordering and inventory holding?
2. Under stochastic demand and supply disruption which inventory management strategy is better, centralisation or decentralisation?
3. How does the inventory management strategy change with respect to varying levels of disruption length and frequency?

## Discrete-Event Simulation Model

A discrete-event simulation model was developed using Arena for Windows version 14 (Fig. 1). Inventory replenishment is considered to follow the base-stock approach and the simulation is run for multiple identical retailers. The model has ten identical retailers with individual demand of 30 per day and standard deviation of 5. Since the retailers are identical, correlation is assumed to be 1, hence the pooled demand comes out to be 300 and pooled standard deviation comes out to be 50. The model can be broken down into three modules, one each for the retailers, the warehouse and the plant. The retailers and the warehouse are assumed to incur backordering and inventory holding cost whereas the plant is assumed to have infinite capacity and can thus fulfill the demands of warehouse without any backordering or holding inventory. All the retailers are assumed to be identical and a cumulative analysis for all the retailers is done (Atilok *et al*., 2010). The base time unit is taken as one day and the model has been replicated for 500 days with 30 replications for each day.

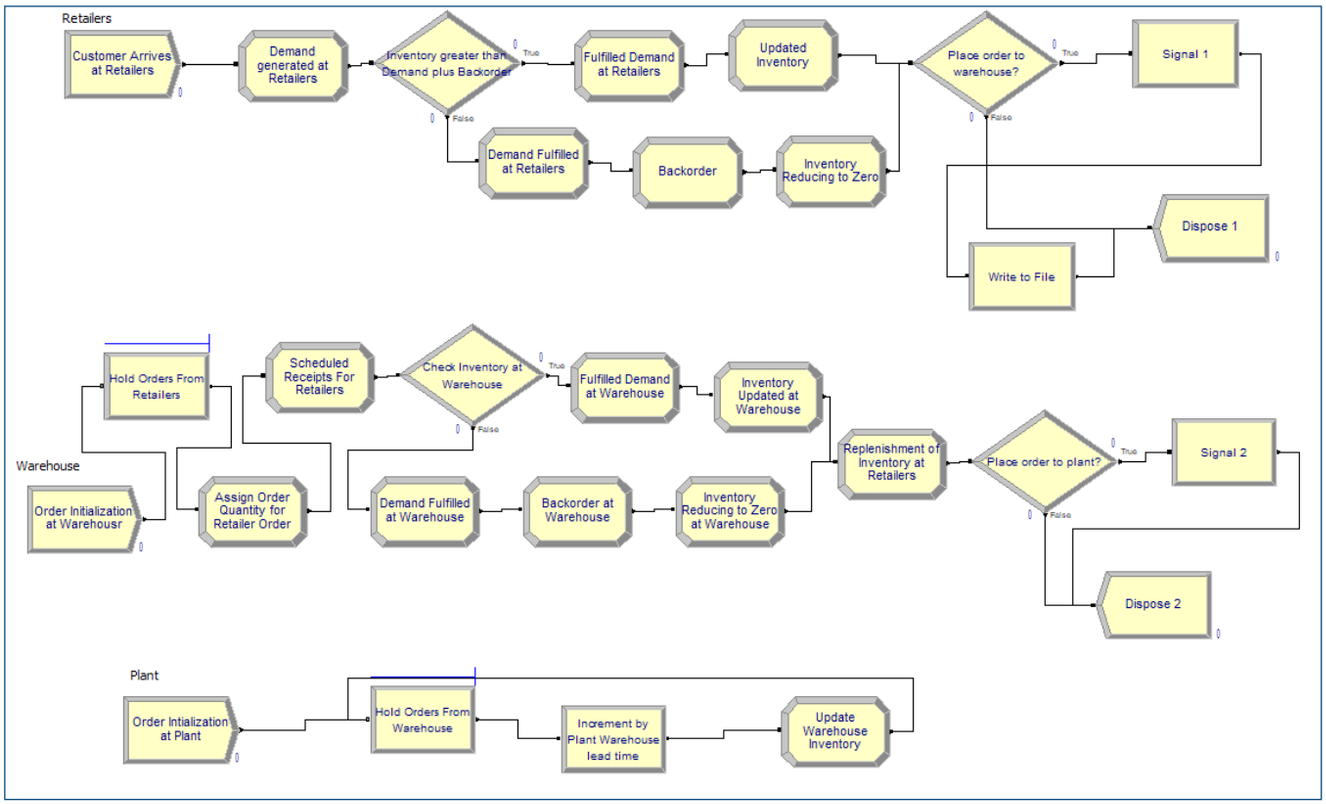



**Fig. 1: Arena Simulation Model Developed in the Study**

## Retailers

Customer demand is generated at the retailers, which is distributed normally with a mean of 300 units per day and a standard deviation of 50 units. The model then computes the inventory level of retailers at a cumulative level and proceeds to calculate demand fulfilled and total backorders. Updated inventory level and inventory position is calculated to compare with the target stock and finally the requisite order is placed to the warehouse. The formula used to calculate inventory position and order quantity is:

$$Q_{ir} = T_{ir} - (OH_{ir} + SR_{ir} - BO_{ir}) \quad \ldots(1)$$

where,

$Q_{ir}$: Quantity ordered by the retailers

$T_{ir}$: Target Stock at the retailers

$OH_{ir}$: On-hand Inventory at the retailers

$SR_{ir}$: Scheduled receipts at the retailers

$BO_{ir}$: Backorders at the retailers

## Warehouse Modeling

The orders received from the retailers act as the input demand for the warehouse. Similar to the retailers' model, the inventory level at warehouse is compared with demand to calculate orders that can be fulfilled and backorders if any. The updated inventory position is compared with the target stock to calculate orders to be placed at the plant as per this formula:

$$Q_{iw} = T_{iw} - (OH_{iw} + SR_{iw} - BO_{iw}) \quad \ldots1$$

where,

$Q_{iw}$: Quantity ordered by the warehouse

$T_{iw}$: Target Stock at the warehouse

$OH_{iw}$: On-hand Inventory at the warehouse

$SR_{iw}$: Scheduled receipts at the warehouse

$BO_{iw}$: Backorders at the warehouse

Further, the orders processed by the warehouse for the retailers are dispatched and delayed as per the lead time which is assumed to be 1 day. Also, the supply disruptions are modelled in the delivery to retailers by disrupting the delivery with additional lead time as done by Schmitt *et al* (2012).

### Plant Modelling

The orders from the warehouse are processed in this module. Since the plant is assumed to infinite capacity, each order received from warehouse is directly processed and delivered with a lead time on 1 day. There is no backordering and inventory holding at plant and no disruptions are introduced in this module for delivery to the warehouse.

## Parameters for Simulation

The model explained in the previous section is varied for parameters to simulate the two inventory policies of centralisation and decentralisation.

### Centralisation

The inventory policy of centralisation entails that the inventory is primarily stocked at the warehouse and retailers maintain only a reasonable level of stock. In this study, we have assumed target stock of 300 units for retailers, which is same as the mean demand per day, the retailers. The warehouse is maintained at target stock of 1000 units that is more than three times the mean demand at retailers.

### Decentralisation

In decentralisation, the inventory is stocked primarily at the retailers as opposed to centralisation. For this model, we have assumed target stock of 1000 units for the retailers and 300 units for the warehouse. Thus, this stocking approach is the reverse of that followed in centralisation. The list of parameters for the complete model and the inventory policies of centralisation and decentralisation are given in Table 1.

**Table 1: List of Parameters Used in Simulation Model**

| *Parameters* | | *Centralisation* | *Decentralisation* |
|---|---|---|---|
| Demand at Retailers (units per day) | | 300 | 300 |
| Standard Deviation (units per day) | | 50 | 50 |
| Target Stock | Retailers (units) | 300 | 1000 |
| | Warehouse (units) | 1000 | 300 |
| Delivery Lead Time | Warehouse to Retailers (days) | 1 | 1 |
| | Plant to Warehouse (days) | 1 | 1 |
| Backordering Cost (per unit) | | 1 | 1 |
| Inventory Holding Cost (per unit) | | 1 | 1 |

## Experimental Design

The type of disruption considered for this model is the stochastic lead-time for delivery to the retailers. Since the model considered is a two-echelon system with one warehouse and multiple retailers, disruptions have been introduced at warehouse with respect to factors:

1. Frequency of disruptions
2. Length of Disruption

As explained in the model, the disruptions are introduced in the warehouse model for the delivery of orders to the retailers. The frequency of disruption is varied from 10% till 90% intervals of 10%. The length of the disruption is varied from 1 day to 10 days. The base model considers the

lead time as 1 day for delivery from warehouse to retailers, so with disruptions the lead times are effectively varied from 2 days to 11 days. Since the disruption frequency is governed by module based on chance, it is possible that disruptions length are duplicated, that is, if another disruption occurs during an already occurring disruption the total disruption length might be distorted. However, since the iterations are run for 500 days such variations tend to even or average out, where the disruption length is only varied from 1 day to 10 days.

## Cost Analysis of Centralised and Decentralised System

The model records inventory and the backorder levels at the retailers as well as the warehouse. These are the two major areas of cost which an operations manager has to deal with. It is quite obvious that if inventory holding cost is given preference the backorder increases and vice versa. The idea of this study is to give an operations manager optimal strategy that needs to be followed in case of supply disruptions which will minimise the overall cost (inventory holding + backordering cost). Parameters chosen to identify a suitable policy under stochastic demand and supply disruptions are cost under different levels of disruption length and frequency. The two cost components considered for comparing centralised and decentralised systems are:

1. Backordering cost
2. Inventory holding cost

Backordering cost is calculated by multiplying average backorder quantity with a constant assumed to be 1 unit of cost in this case. Similarly, inventory holding cost is calculated by multiplying average inventory with a constant, again assumed to 1 unit. Moreover, this study compares the total cost for the two systems to identify a suitable policy in addition to individual cost components of backordering and inventory holding costs. Further, this study tries to identify the main cost component for the total cost and the individual characteristics of cost components. The final output that the model will produce will help an operations manager to come up with an optimal inventory management policy in case of supply disruptions (Mak & Shen, 2012).

## Cost Analysis for the Centralised System

In this section, we examine the results obtained from simulations of the centralised system. Figs. 2 and 3 respectively present the backordering and inventory holding costs for a centralised system under varied disruption length and frequency. As it is evident from this, both costs are lower with lower levels of disruptions. For disruption frequencies less than 20%-30% and disruption length less than about 2-4 days the costs are lower than those with higher orders of disruption.


**Fig. 2: Plot of Backordering Costs for Centralised System Under Varied Disruption Length and Frequency**

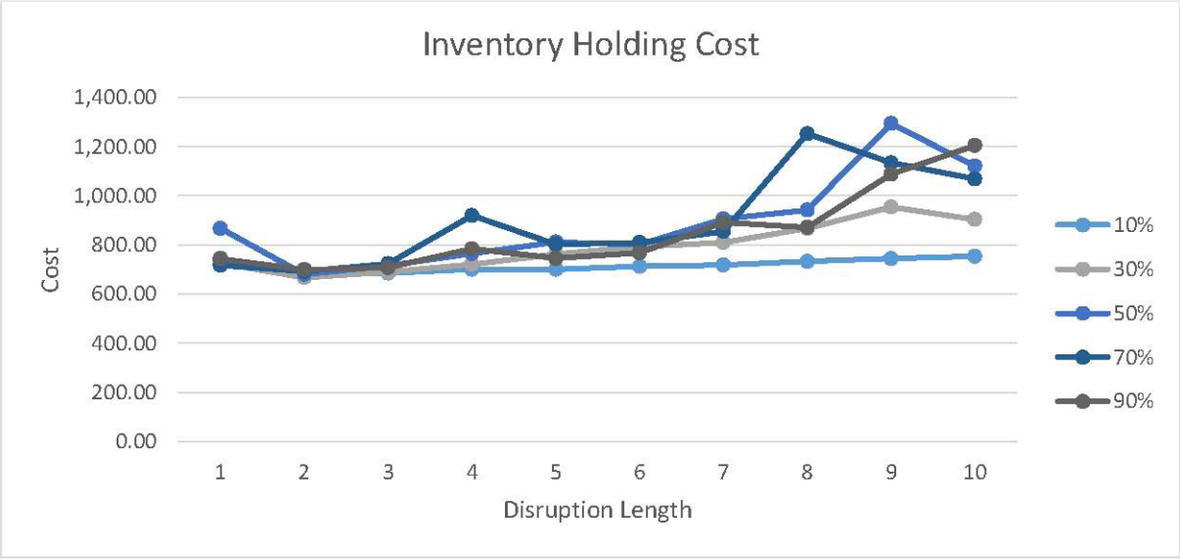



**Fig. 3: Plot of Inventory Holding Costs for a Centralised System Under Varied Disruption Length and Frequency**

## Cost Analysis for the Decentralised System

This section presents the results obtained from simulations of the decentralised system. Figs. 4 and 5 respectively present the backordering and inventory holding costs for a decentralised system under varied disruption length and frequency. Backordering costs follow the same pattern as in centralised system. Costs are lower for low levels of disruption with disruption frequency lesser than 10%-20% and disruption length lesser than 2-3 days. However, the inventory holding costs follow an opposite pattern to that of the centralised system, since they are observed to be lower for higher levels of disruption. Inventory holding costs are decreasing as the level of disruption increases and within disruption frequency of 80%-90% the costs are lowest.

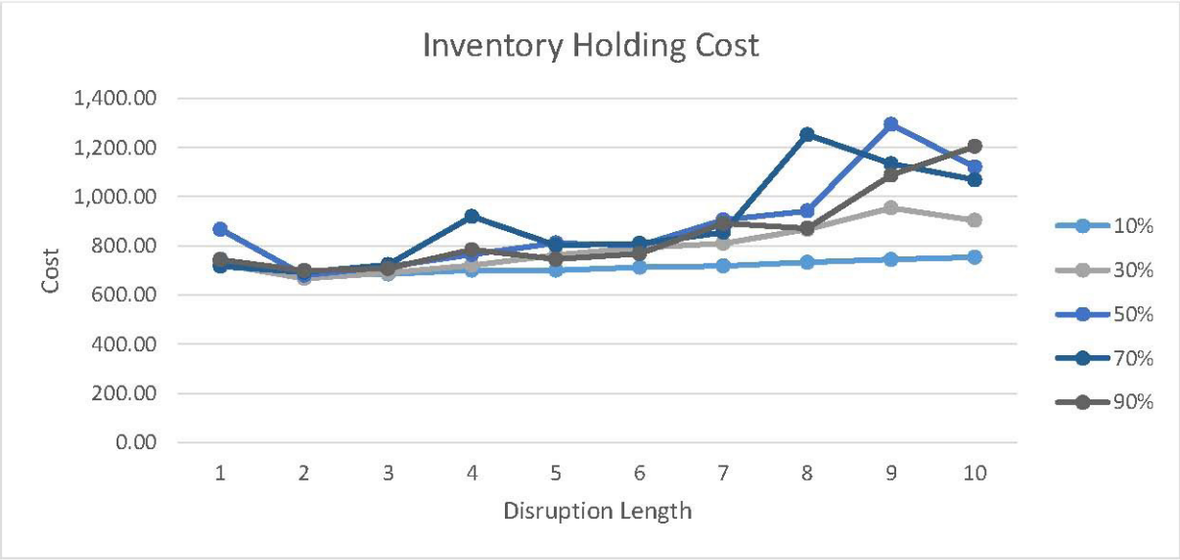



**Fig. 4: Plot of Backordering Costs for De-Centralised System Under Varied Disruption Length and Frequency (percentages are for disruption frequencies and disruption length is lead time delay)**

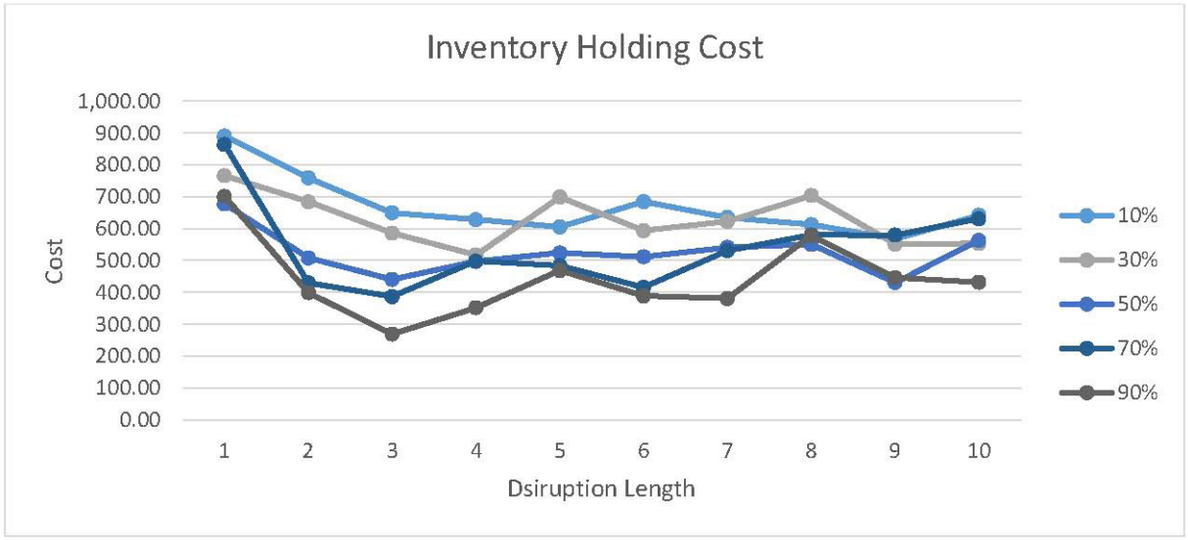



**Fig. 5: Plot of Inventory Holding Costs for De-Centralised System Under Varied Disruption Length and Frequency (percentages are for disruption frequencies and disruption length is lead time delay)**

## Cost Comparisons Between Centralised and Decentralised Systems

In this section, we compare the backordering, inventory holding and total costs between centralised and decentralised systems. We do the comparison by plotting the ratios of each cost under centralisation to that under decentralisation (see Figs. 6-8 respectively for backorder, holding and total cost ratios).

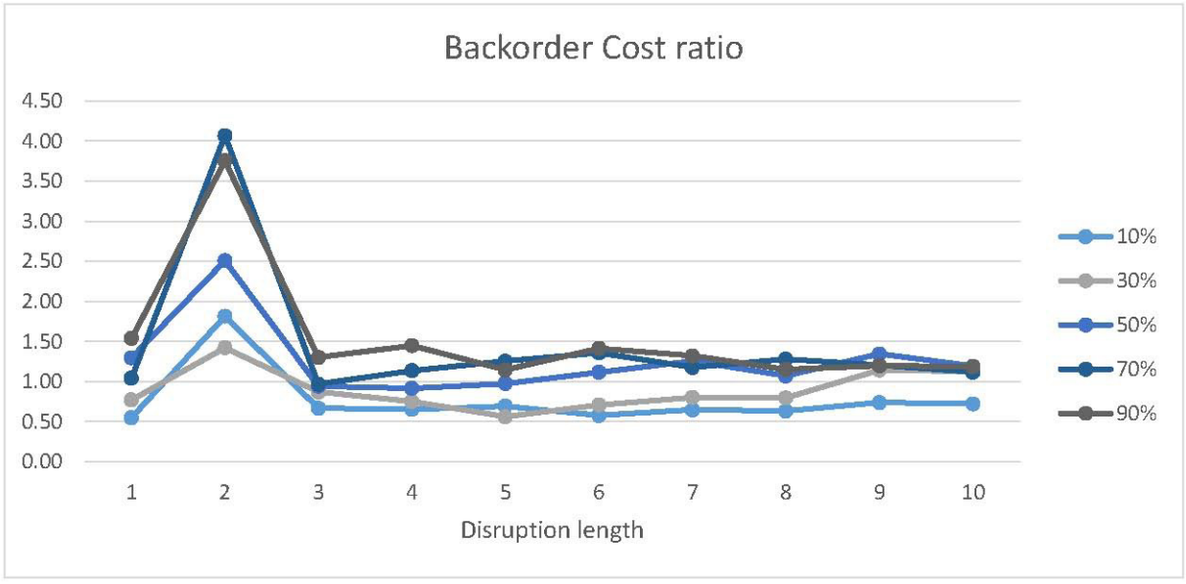



**Fig. 6: Plot of Backorder Cost Ratio Between Centralisation and De-Centralisation Under Varied Disruption Length and Frequency**

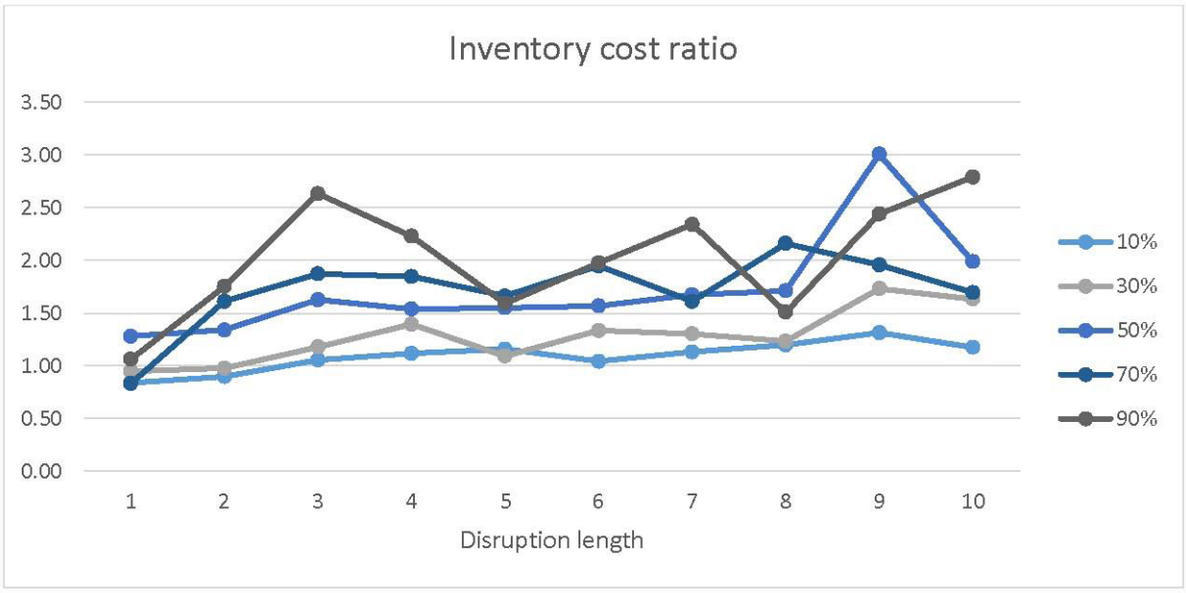



**Fig. 7: Plot of Holding Cost Ratio Between Centralisation and De-Centralisation Under Varied Disruption Length and Frequency**

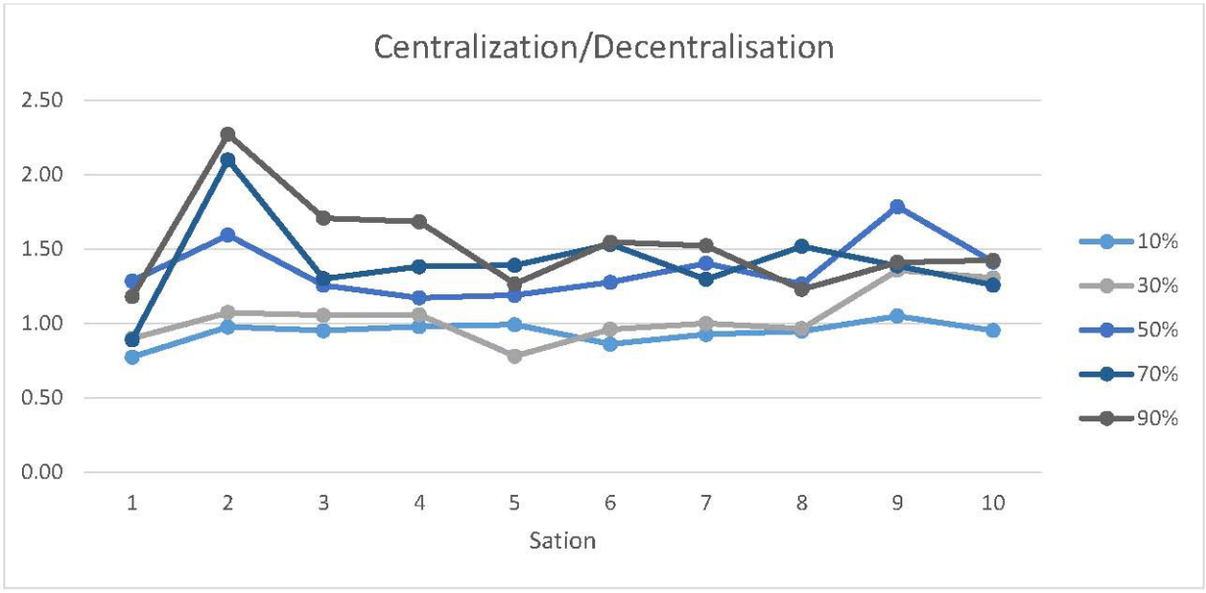



**Fig. 8: Plot of Total Cost Ratio Between Centralisation and De-Centralisation Under Varied Disruption Length and Frequency (percentages are for disruption frequencies and disruption length is lead time delay)**

It is noted that the backordering costs for the centralised system are lower for almost half of the cases. Lower costs for centralisation go up to 50% of disruption frequency. Therefore, decentralised system has lower backorder costs for only higher levels of disruption. Unlike backorder costs, holding costs for centralised system are lower only for 10% of cases. This is concentrated for lower levels of disruption length (1-2 days) and lower levels of disruption frequency (40%). If we consider only holding costs under disruption, then decentralised system will be a better policy for minimising costs. Finally, when compared the total costs, the centralised system has lower values than the decentralised system for almost one-third of cases. Mostly, the centralised system fares well under

lower levels of disruption, that is, when either or both the disruption frequency is low (10%-20%) or disruption length is low (1-2 days). Thus, a decentralised system is more likely to yield lower costs under higher disruption levels and if we are not expecting disruptions or lower level of disruption, centralisation may be a better policy.

## Results

Summarising the analyses presented above, it is noted that:

- For total of holding and backordering costs, the centralisation approach is a better strategy when disruption levels and disruption are lower, while decentralisation is better as disruption increases.
- From the perspective of backordering costs alone, centralisation is better strategy for lower levels of disruption length and frequency, but the pattern was slightly different than total cost because in relatively more number of cases centralisation gave minimum cost.
- Holding costs also moved in the same pattern as the combined cost and it is noted that the overall combined cost is majorly governed by holding cost. In this case too, centralisation was a better strategy for lower levels of disruption frequency and length.

Holding cost is the governing factor for the combined costs for both the systems. While, backorder cost has almost equal instances where the centralised and decentralised systems have lower costs, inventory costs are highly skewed towards decentralised system. Overall, holding costs are the reason for bringing down the number of instances where centralised system has a lower cost. Clearly, under low or no disruptions centralised inventory policy would fare better. However, if the disruptions are frequent and large, decentralisation must be adopted to bring down the cost and ensure delivery.

## Conclusion and Future Research Directions

Firm survival in the modern business environment is no longer an issue of one firm competing against another but has, instead, become an issue of one supply chain competing against another supply chain. In most of the top global firms supply chain disruptions and their associated operational and financial risks are the most pressing concerns (Green, 2004).

Indeed, research on issues ranging from business continuity planning (e.g., Zsidisin, Melnyk, & Ragatz, 2005) to supply chain vulnerability (e.g., Svensson, 2000) to supply chain resilience (e.g., Sheffi and Rice, 2005) to supply chain risks (e.g., Chopra & Sodhi, 2004) has not only confirmed the costliness of supply chain disruptions but has also contributed insights to this very concern. Our study research has provided additional value to the rich and growing body of knowledge on supply disruptions, particularly on the issue of choosing between centralisation and decentralisation. Some research has also discussed supply chain cost under production disruption when retailers compete with price and service levels (Giri & Sarker, 2016). This study mostly focuses on decentralised system of inventory management. To add to this, our study further focuses on both centralised and decentralised inventory management policies under supply disruptions.

In our study we have analysed holding and backorder costs under stochastic demand and supply disruptions for the two-inventory policies, centralisation and decentralisation. In the course of this, we have made certain assumptions. The unit cost for both inventory holding and backorder are assumed to be one unit. For the sake of simplicity, all the retailers are assumed to be similar. However, this might not be possible in real scenario and a differentiation of retailers may lead to different results. Costs for holding unit inventory and backorder are assumed to be one unit, which might have an impact on the level of costs, observed for the two policies. If the costs for holding inventory are different for warehouse and retailers, then this might yield different results. Disruptions are introduced in supply link to the retailers. Disruptions could also be introduced at plant to warehouse supply.

In this study, we have also assumed 10 identical retailers. However, it is possible that the results may vary if the number of retailers is varies. The same could be analysed further in future course of study. Future studies can also explore relaxing one or more of our assumptions and take the discussion forward.

## References

Al-Rifai, M., & Rossetti, M. (2007). An efficient heuristic optimization algorithm for a two-echelon (R, Q) inventory system. *International Journal of Production*

*Economics*, 109(1-2), 195-213. Retrieved from http://dx.doi.org/10.1016/j.ijpe.2006.12.052

Altiok, T., & Melamed, B. (2007). *Simulation modeling and analysis with Arena* (1st ed.). Amsterdam: Academic Press.

Arreola-Risa, A., & DeCroix, G. (1998). Inventory management under random supply disruptions and partial backorders. *Naval Research Logistics*, 45(7), 687-703. Retrieved from http://dx.doi.org/10.1002/(sici)1520-6750(199810)45:7<687::aid-nav3>3.3.co;2-#

Atan, Z., & Snyder, L. (2012). Disruptions in one-warehouse multiple-retailer systems. *SSRN Electronic Journal*. Retrieved from http://dx.doi.org/10.2139/ssrn.2171214

Bollapragada, R., Rao, U. S., & Zhang, J. (2004). Managing inventory and supply performance in assembly systems with random supply capacity and demand. *Management Science*, 50(12), 1729-1743.

Boute, R. N., Disney, S. M., Lambrecht, M. R., & Van Houdt, B. (2009). Designing replenishment rules in a two-echelon supply chain with a flexible or an inflexible capacity strategy. *International Journal of Production Economics*, 119, 187-198.

Bulut, Z., & Snyder, L. V. (2009). Supply disruptions in a one-warehouse multiple-retailer system, Working Paper, P.C. Rossin College of Engineering and Applied Sciences, Lehigh University, Bethlehem, PA.

Chopra, S., & Sodhi, M. S. (2004). Managing risk to avoid supply-chain breakdown. *MIT Sloan Management Review,* 46(1), 53-61.

Deleris, L. A., & Erhun, F. (2005). *Risk management in supply networks using Monte-carlo simulation,* in Kuhl, M.E., Steiger, N.M., Armstrong, F.B, Joines, J.A. (Eds), Proceedings of the 2005 Winter Simulation Conference, pp. 1643-49.

Giri, B., & Sarker, B. (2015). Coordinating a two-echelon supply chain under production disruption when retailers compete with price and service level. *Operational Research*, 16(1), 71-88. Retrieved from http://dx.doi.org/10.1007/s12351-015-0187-8

Green, M. (2004). Loss/risk management notes: Survey: Executives rank fire, disruptions top threats. *Best's Review,* 105(5), 105.

Hopp, W. J., & Liu, Z. (2006). Protecting supply chain networks against catastrophic failure, Working Paper, Dept. of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

Mak, H., & Shen, Z. (2012). Risk diversification and risk pooling in supply chain design. *IIE Transactions*, 44(8), 603-621. Retrieved from http://dx.doi.org/10.1080/0740817x.2011.635178

Qi, L., Shen, Z. J. M., & Snyder, L. V. (2010). The effect of supply disruptions on supply chain design decisions. *Transportation Science*, 44(2), 274-289.

Rong, Y., Atan, Z., & Snyder, L. V. (2015). Heuristics for base-stock levels in multi-echelon distribution networks. Working Paper, Lehigh University, Bethlehem, PA.

Schmitt, A. J. (2011). Strategies for customer service level protection under multi-echelon supply chain disruption risk. *Transportation Research Part B*, 45, 1266-1283.

Schmitt, A. J., & Singh, M. (2009). *Quantifying supply chain disruption risk using Monte carlo and discrete-event simulation*, In Rossetti, M.D., Hill, R.R., Johansson, B., Dunkin, A., Ingalls, R.G. (Eds.), Proceedings of the 2009 Winter Simulation Conference, pp. 1237-1248.

Schmitt, A., & Singh, M. (2012). A quantitative analysis of disruption risk in a multi-echelon supply chain. *International Journal of Production Economics*, 139(1), 22-32. Retrieved from http://dx.doi.org/10.1016/j.ijpe.2012.01.004

Schmitt, A., Snyder, L., & Shen, Z. (2010). Inventory systems with stochastic demand and supply: Properties and approximations. *European Journal of Operational Research*, 206(2), 313-328. Retrieved from http://dx.doi.org/10.1016/j.ejor.2010.02.029

Schmitt, A., Sun, S., Snyder, L., & Shen, Z. (2014). Centralization versus decentralization: Risk pooling, risk diversification, and supply chain disruptions. *Omega*, 52, 201-212. Retrieved from http://dx.doi.org/10.1016/j.omega.2014.06.002

Sheffi, Y. (2005). *The resilient enterprise: Overcoming vulnerability for competitive advantage*. MIT Press, Cambridge, MA.

Sheffi, Y., & Rice, J. (2005). A supply chain view of the resilient enterprise. *MIT Sloan Management Review*, 47(1), 41-48.

Sheffi, Y., (2001). Supply chain management under the threat of international terrorism. *The International Journal of Logistic Management*, 12(2), 1-11.

Simchi-Levi D., Snyder, L., & Watson, M. (2002). Strategies for uncertain times. *Supply Chain Management Review*, 6(1), 11-12.

Snyder, L. V., Atan, Z., Peng, P., Rong, Y., Schmitt, A, J., & Sinsoysal, B. (2016). OR/MS Models for supply chain disruptions: A review. *IIE Transactions*, 48(2), 89-109.

Svensson, G. (2000). A conceptual framework for the analysis of vulnerability in supply chains. *International Journal of Physical Distribution & Logistics Management*, 30(9), 731-750. Retrieved from http://dx.doi.org/10.1108/09600030010351444

Tang, C. S. (2006). Perspectives in supply chain risk management. *International Journal of Production Economics*, 116, 12-27.

Tang, C. S., & Tomlin, B. (2008). The Power of flexibility for mitigating supply chain risks. *International Journal of Production Economics*, 116, 12-27.

Tee, Y., & Rossetti, M. (2002). A robustness study of a multi-echelon inventory model via simulation. *International Journal of Production Economics*, 80(3), 265-277. Retrieved from http://dx.doi.org/10.1016/s0925-5273(02)00259-1

Zsidisin, G., Melnyk, S., & Ragatz, G. (2005). An institutional theory perspective of business continuity planning for purchasing and supply management. *International Journal of Production Research*, 43(16), 3401-3420. Retrieved from http://dx.doi.org/10.1080/00207540500095613

# Guidelines for Authors

International Journal of Business Analytics and Intelligence welcomes original manuscripts from academic researchers and business practitioners on the topics related to descriptive, predictive and prescriptive analytics in business. The authors are also encouraged to submit perspectives and commentaries on business analytics, cases on managerial applications of analytics, book reviews, published-research paper reviews and analytics software reviews based on below mentioned guidelines:

Journal follows online submission for peer review process. Authors are required to submit manuscript online at http://manuscript.publishingindia.com

**Title:** Title should not exceed more than 12 Words

**Abstract:** The abstract should be limited to 150 to 250 words. It should state research objective(s), research methods used, findings, managerial implications and original contribution to the existing body of knowledge

**Keywords:** Includes 4–8 primary keywords which represent the topic of the manuscript

**Main Text:** Text should be within 4000-7000 words Authors' identifying information should not appear anywhere within the main document file. Please do not add any headers/footers on each page except page number. Headings should be text only (not numbered).

**Primary Heading:** Centered, capitalized, and italicized.

**Seconday Heading:** Left justified with title-style capitalization (first letter of each word) and italicized.

**Tertiary Heading:** Left justified and indented with sentence-style capitalization (first word only) in italics.

**Equations:** Equations should be centered on the page. If equations are numbered, type the number in parentheses flush with the left margin. Please avoid using Equation Editor for simple in-line mathematical copy, symbols, and equations. Type these in Word instead, using the "Symbol" function when necessary.

**References:** References begin on a separate page at the end of paper and arranged alphabetically by the first author's last name. Only references cited within the text are included. The list should include only work the author/s has cited. The authors should strictly follow APA style developed by American Psychological Association available at American Psychological Association. (2009). Publication manual of the American Psychological Association (6th Ed.). Washington, DC.

## Style Check

To make the copyediting process more efficient, we ask that you please make sure your manuscript conforms to the following style points:

Make sure the text throughout the paper is 12-point font, double-spaced. This also applies to references.

Do not italicize equations, Greek characters, R-square, and so forth. Italics are only used on p-values.

Do not use Equation Editor for simple math functions, Greek characters, etc. Instead, use the Symbol font for special characters.

Place tables and figures within the text with titles above the tables and figures. Do not place them sequentially at the end of the text. Tables and figures must also be provided in their original format.

Use of footnotes is not allowed; please include all information in the body of the text.

## Permissions

Prior to article submission, authors should obtain all permissions to use any content if it is not originally created by them.

When reproducing tables, figures or excerpts from another source, it is expected to obtain the necessary written permission in advance from any third party owners of copyright for the use in print and electronic formats. Authors should not assume that any content which is freely available on the web is free to use. Website should be checked for details of copyright holder(s) to seek permission for resuing the web content

## Review Process

Each submitted manuscript is reviewed first by the chief editor and, if it is found relevant to the scope of the journal, editor sends it two independent referees for double blind peer review process. After review, the manuscript will be sent back to authors for minor or major revisions. The final decision about publication of manuscript will be a collective decision based on the recommendations of reviewers and editorial board members

## Online Submission Process

Journal follows online submission for peer review process. Authors are required to register themselves at http:// manuscript. publishingindia.com prior to submitting the manuscript. This will help authors in keeping track of their submitted research work. Steps for submission is as follows:

1. Log-on to above mentioned URL and register yourself with "International Journal of Business Analytics & Information"
2. Do remember to select yourself as "Author" at the bottom of registration page before submitting.
3. Once registered, log on with your selected Username and Password.
4. Click "New submission" from your account and follow the 5 step submission process.
5. Main document will be uploaded at step 2. Author and Co-author(s) names and affiliation can be mentioned at step 3. Any other file can be uploaded at step 4 of submission process.

**Editorial Contact**
Dr. Tuhin Chattopadhyay
Email: dr.tuhin.chattopadhyay@gmail.com
Ring: 91-9250674214

**Online Manuscript Submission Contact**
Puneet Rawal
Email: puneet@publishingindia.com
Ring: 91-9899775880

# International Journal of Business Analytics and Intelligence

## SUBSCRIPTION DETAILS

Dispatch Address:-
The Manager,
International Journal of Business Analytics and Intelligence
Plot No-56, 1st Floor
Deepali Enclave, Pitampura
New Delhi -110034
Ph - 9899775880

### Subscription Amount for Year 2018

| | Print | Print + Online |
|---|---|---|
| Indian Region | Rs 2700 | Rs 3400 |
| International | USD 150 | USD 180 |

Price mentioned is for Academic Institutions & Individual. Pricing for Corporate available on request. Price is Subject to change without prior notice.

Payment can be made through D.D./at par cheque in favour of "Publishing India Group" payable at New Delhi and send to above mentioned address.

## Disclaimer

The views expressed in the Journal are of authors. Publisher, Editor or Editorial Team cannot be held responsible for errors or any consequences arising from the use of Information contained herein. While care has been taken to ensure the authenticity of the published material, still publisher accepts no responsibility for their inaccuracy.

Journal Printed at Anvi Composers, Paschim Vihar.

## Copyright


Printed and published by Publishing India Group, New Delhi. Any views, comments or suggestions can be addressed to – Coordinator, IJBAI, info@publishingindia.com


**www.manuscript.publishingindia.com**

PublishingIndia.com

Publishing India Group

Plot No. 56, 1st Floor, Deepali Enclave
Pitampura, New Delhi-110034, India
Tel.: 011-47044510, 011-28082485
Email: info@publishingindia.com
Website: www.publishingindia.com

ISSN 2321-1857
02
9 772321 185001