Rising Business Manager Prediction in Fixed Line Operators Using Classification Models

Submitted By
Naveed-ul-Islam
818-FBAS/MSCS/F14

Supervisor
Dr. Ali Daud

Department of Computer Science & Software Engineering
Faculty of Basic and Applied Sciences
International Islamic University Islamabad
2016

1. Data mining

رَّبِّ زِدْنِیْ عِلْمًا ۝

(طٰهٰ: 114)

اے میرے رب! میرے علم میں اضافہ فرما۔

In the name of

Allah,

The most Merciful and Compassionate the most Gracious and the Beneficent whose help and Guidance we always solicit at every step, and every moment.

DECLARATION

I **Naveed ul Islam s/o Abdul Rauf**

Registration No. **818-FBAS/MSCS/F14**

Student of MS in Computer Science at International Islamic University do hereby solemnly declared that the thesis entitled "**Rising Business Manager Prediction for Fixed Line Operators Using Classification Models**", submitted by me in partial fulfilment of MS degree in Computer Science, is my original work, except where otherwise acknowledged in the text, and has been submitted or published earlier and shall not, in future, be submitted by me for obtaining any degree from this or any other university or institution.

Naveed ul Islam

818-FBAS/MSCS/F14

ACKNOWLEDGEMENT

All praise to Almighty Allah who bestowed me life and enabled me to reveal the insights that were necessary for accomplishment of this major career milestone. I am really grateful to my research supervisor Dr. Ali Daud and indeed with his guidance and encouragement this objective has been achieved. I also extend my gratitude to my parents and family for supporting me in my life and throughout the career.

I further admire the support of my friends **Muhammad Siddique Ahmed, Safeer Ahmed, Farrukh Javed, Shakeel Hussain, Malik Khizer Hayat** and **Khalid Ansar** in helping me on different steps of the project.

Naveed ul Islam

818-FBAS/MSCS/F14

LIST OF FIGURES

LIST OF TABLES

Abstract

The business manager is a core value function for the telecom industry and the classification models are required to predict the rising business manager so that an intelligent business strategy can be devised. Unlike the most common classification models that only relies on history of achievements to determine the rising behavior, we evaluated the performance of co-business managers to predict the rising business manager using machine learning techniques.

We formulated fifteen features of co-business manager and his top management classified into two different scenarios of evaluation. One scenario conceiving average revenue (AR) and other having average relative increase in revenue (ARIR) as class label. All instances of both scenarios are sorted randomly into different size of data sets (10, 20, 30 to 100) drilled through four classifiers (NN, SVM, NB and BN) using 5-fold cross validation.

We further proposed three different methods against baseline results of co-business manager to gauge the effectiveness. The feature, category and model based experimental results derived after feature subset selection in terms of precision, recall and f-measure depicts that discriminative models particularly SVM has shown better results because the features are not linearly separable and needs vector support. We also ranked (top and bottom 10) business managers and compared the results with business scores that are derived based on value of feature for the business.

Department of Computer Science and Software Engineering

International Islamic University Islamabad, Pakistan

Date: Oct 28, 2016

Final Approval

This is to certify that we have read and evaluated this thesis titled "**Rising Business Manager Prediction in Fixed Line Operators Using Classification Models**" submitted by **Naveed-ul-Islam** bearing registration No. **818-FBAS/MSCS/F14.** It is our judgment that this thesis is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the award of degree of **Master of Science in Computer Science.**

Committee

External Examiner
Dr. Iftikhar Azeem
Assistant Professor
Department of Computer Science
COMSATS Institute of IT, Islamabad

Internal Examiner
Dr. Jamal Abdul Nasir
Assistant Professor
Department of Computer Science & Software
International Islamic University, Islamabad

Research Supervisor
for **Dr. Ali Daud** (on Leave)
Assistant Professor
Department of Computer Science & Software
International Islamic University, Islamabad

for

CHAPTER 1

INTRODCUTION

1. Introduction

The growing need of every business is valuable insights and recommendations derived on the top of information generated by their systems on each activity. The trend analysis, statistical functions, projection and prediction are effective tools to stimulate business ahead of time for in-time decisions and planning. In specific, the decision to map and place HR particularly managers on right place and in a right customer segment is quite vibrant.

We addressed the main challenge of identifying the rising business manager based on performance of co-business managers and team rather to rely on manager's profile and history. We evaluated two different scenarios of average revenue and average relative increase in revenue against fifteen core features of fixed line operator and explored them using machine learning techniques. We proposed three different methods against baseline and have observed that proposed methods performs better. We further compared the rising business manager results with business rating to gauge effectiveness of this exercise.

The challenge of hypothetical determination the persuasive literature based on paper citation and trends to propose a Future Influence Prediction FIP indicator after a certain time has also been resolved. Many regression models have been applied such as k-Nearest Neighbor (kNN), Linear Regression (LR), support vector regression (SVR), CART and Gaussian predication process (GPR) to learn and evaluate accuracy based on determination coefficient. A significant improvement in average of mean precision has been observed. The content features (novelty, ranking of the topic, diversity), feature of the authors (Rank of the Author, H-index, author influence history, productivity, sociality, authority and versatility), venue features and temporal features are evaluated through academic research database ArnetMiner to find FIP [1].

1.1 Rising star

The rising stars are emerging experts that currently carries relatively low profiles but can have significant/exceptional role in business disciplines / areas [2]/ [3]. In essence the business success factors in fixed line operators are based on productivity of business managers. The objective is to identify the new manager as rising based on certain aligned features of co-business managers and their team to predict rising business manager.

1.2 Rising vs. Expert Business Manager

The expert business managers are outstanding managers, who delivers and leads by example and are ranked on top based on management defined goals and key performance indicators (KPI). The business manager leads the particular business area by driving his/her team with a common goal of efficient and productive business management. A new business manager is rising or emerging, if the co-business managers and his/her team delivers outstandingly.

1.3 Prediction Using Classification Model

The data mining in nut shell is a collection of techniques that are required for the extraction of relevant knowledge from data. The classification and prediction is one of the tasks of data mining. It's based on supervised learning that classifies the data items into predefined class labels [4]

Every classification algorithm progresses on a training data set that comprises of attributes and outcome. The algorithm attempts to ascertain the relationship among the attributes that will support the prediction of the outcome. On discovery of relationship among attributes, an un-seen / un-interpreted data set called prediction set is given to the algorithm [5].

The classification models built from an input data set are used to predict the future data trends. The decision tree, Naïve Bayes classifiers and SVM are three of the main algorithms for data classification.

The decision tree generates a tree from the given data using simple equations based on the calculation of gain ratio. The gain ratio assigns weights to the attributes used so that the researcher can identify the most operative attributes on the foreseen target. Resultantly, a decision tree would be built with classification rules generated from it [6].

The Naive Bayes (NB) classifier is also a classification technique used to predict a target class. It's based on the calculations of probabilities, namely Bayesian theorem. Its classifiers are more precise and real, and are more sensitive to new data added to the dataset but its compute intensive [6]. The NB is a probabilistic classification method that relates naïve hypothesis with Bayes algorithm for every pair of features. It can handle both continuous and categorical independent variables and adopts that features are statistically independent [7].

The support vector machine (SVM) is also a supervised learning model / algorithm that analyzes the data and recognizes the patterns, required for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier [8].

MEMM (Maximum entropy markov model) or CMM (Conditional Markov Model) encapsulates the functionality of both HMM and maximum entropy for labeling of sequential data. In nut shell, it extends the MEC (maximum entropy classifier) with features. The unknown parameters are assumed to be connected in a markov chain rather than independent to one another [7].

1.4 Fixed line operators

In fixed line the signal is transmitted through a physical line, wire and pole unlike wireless network where signal is transmitted in air. The fixed line or land line is comprising of metal wire telephone line. The term landline is also used to describe a connection between two or more points that consists of a dedicated physical cable [9].

CHAPTER 2

LITERATURE REVIEW

2. Literature Review

The individuals that are currently not stars or not experts but have latent to be ranked on top in any specific field are rising stars [2]. In academic social networks, the features of co-authors and co-citations have been measured to rank researchers and publications based on quantity and quality of work [7].

A trivial set of energies have so far been engaged to determine the rising scholars, researchers or players based on performance of co-members. It has further been proposed an innovative technique Coca Rank for ranking of citation count in academic networks to find rising stars. In addition, right time span for prediction of rising stars has also been determined with a prime focus on researcher's social interaction along with scientific potential through a meter Collaboration Caliber. The rising stars have also been defined in terms of researchers as individuals if compared with colleagues, are not currently outstanding but have tendency to grow and become scholars. [10]

The IT professionals categorized the interpersonal and communication skills comparatively more in importance than technical users. The engineer's performance is based on interpersonal skills not on technical background. Further the emphasis is on creative problem-solving and ability to abstract business problems [11].

Forecasting of academic talent by using past experience knowledge discovered from related databases has been performed and then the classification model is used for talent forecasting. The result indicates different challenges associated with good classification model [12]. A paper also predicts the employee talent based on existing and past knowledge of the employee [13].

A model has been proposed with a stance that the quality of team work is very important and it impacts each employee keeping in view the effect of technical and human IT infrastructure capabilities on IT project success, which subsequently affects team commitment that is crucial to IT project success [14].

For competitive edge, it's quite decisive to focus on quality and competency of employees. Unfortunately, the high technology companies are facing high employee turnover rate and it became very hard to hire a right resource. The development of effective data mining based personnel selection framework to hunt a suitable talent is inevitable.

A study carried out to fill the gap by developing a data mining framework based on decision tree and association rules to generate useful rules for personnel selection [15].

Based on empirical study of 178 industry projects, it has been identified that the success of any product depends on the skills and competency of the product manager [16].

In a methodology, the bibliography networks are excavated in unsubstantiated way using clustering (power graphs) based on **co-authors** information, volume of publications and venues of these publications using DBLP databases. In addition, an author evolution paradigm is established on well established, rising, declining authors against publication rate. The formula to calculate evolution index and average relative increase is defined. [17]

The first step of data classification is learning in which a model that defines a determined set of classes is built by evaluating a set of training data cases. Each case is supposed to belong to a predefined class. In second step, the model is tested by using a different dataset to determine the correctness of the model. If the model correctness or accuracy is acceptable, the model can further be used for classification of unknown classes. This model will then act as a classifier in decision making process [4].

The decision tree, Bayesian methods, rules based algorithms and neural networks are the techniques that can be used for classification. The decision tree classifiers are the well-known technique because the assembly of tree does not require any domain expert knowledge. The decision tree can yield a model with rules that are human readable and interpretable. The decision tree has the advantages of easy interpretation for decision makers to compare with their domain knowledge for validation and decision justification [4].

To address over fitting issue of decision tree, a conscious working on two independent hybrid mining algorithms is carried out to improve the classification accuracy rates of decision tree and naive Bayes classifiers for the classification of multi-class problems. Further, it has also been verified the enactments of the two anticipated hybrid algorithms against those of the existing DT and NB classifiers respectively using the classification accuracy, precision, sensitivity-specificity analysis, and 10-fold cross validation on 10 real benchmarks [18].

The Bayesian latent variable model has been used with classification and regression tree approach for a bank credit grant department to predict the applicant's future performance based on current information of credit procedure and reimbursements. The cardholder's credit relic provides the most noteworthy proof in credit scoring [19].

We found explanation of advantages of classification and regression tree (CART) methods as simplicity of results, easy implementation, nonlinear estimation, being non-parametric, accuracy and stable. For knowledge discovery and mining, the most powerful approach is decision trees. Its strength is to explore the large and complex data sets to find useful hidden patterns. The decision tree enriches the model by self-learning [20] /[21].

MEMM (Maximum entropy markov model) or CMM (Conditional Markov Model) encapsulates the properties of both HMM and maximum entropy for marking of progressive data. In nut shell, it extends the maximum entropy classifier with structures. The unidentified constraints are supposed to be linked in a markov chain rather than autonomous. [7]

2.1 Problem Statement

For an intelligent business strategy and decision making in fixed line operators, the problem is to identify a best fit business manager before his actions in the same business. This will ultimately control the decisions of business manager hiring / reallocation and will guarantee the business growth and the competitive edge in the market mix. It will also help the Telco business to use the right resource on right task by prediction not by experiment.

The problem is to predict a new business manager as rising star based on certain aligned features of co-business managers and overall team progress not on the basis of his/her own profile and work history.

Given a series of n drill examples (X1, y1), (X2, y2) . . . (Xn, yn), where n is the total numbers of business managers, Xi is a feature vector of business manager ai, where Xi ∈ Rm, m is total number of features and yi ∈ {-1, +1}. To classify either a business manager ai is a rising star or not, the function for prediction is produced as

$$y = \mathrm{FRS}\ (a / X) \qquad (1)$$

Where,

$$F_{RS}\,(a / X) = \begin{bmatrix} \geq 0 \text{ if } y = +1, \text{ rising star} \\ < 0 \text{ if } y = -1, \text{ not rising star} \end{bmatrix} \qquad (2)$$

The objective is to study a predictive function FRS (.), to predict either a business manager 'a' is a rising star or not after a certain time period Δt.

$$\hat{y} = \widehat{F}\text{RS}\ (a\,/\,\mathbf{X},\ \Delta t) \tag{3}$$

2.2 Research Objectives

1. Collection of data of business managers
2. Evaluate features
3. Defining AR and ARI
4. For finding rising business manager, we apply classification modeling techniques such as Neural Net, Support Vector Machine (SVM), Bayes Net and Naive Bayes.
5. Comparing results of proposed methods from baseline.
6. Predicting top and bottom 10 business managers using rising star score and ARI (Average Relative Increase).

CHAPTER 3

RESEARCH METHODOLOGY

3. Research Methodology

To predict a manager as a rising business manager, the fifteen features of the co-business managers, senior business manager and the team's regional general manager have been formulated. On the basis of weightage and point calculations the new business manager will be determined as rising or not.

Based on trend of last three months' progress, the high limit and low limit for calculation of score is determined.

Achievement Trend (Variance) = [MAX (Current KPI Achieved, Last three months' average KPI achieved)-MIN (Current KPI Achieved, Last three months' average KPI achieved))/ Last three months' average KPI achieved]

The fifteen features that will be used for classification of rising and not rising business manager are grouped into three different categories i.e. Co-BM, SBM and RGM. Co-BM refers to the feature set of co-business managers that includes nine features. SBM refers to the feature set of Senior business manager that includes three features and RGM category covers feature set of regional general manager that also includes three features. The features segregation into each category is produced in table 1.

Table 1: Category Wise Feature Sets

Co-BM	On-Time Provisioning	On-Time Rectification	Faults Registered
	Repeat Faults	Repeat Telephone	Net Adds
	Disconnection	Customer Denied Rectification	Customer Denied Installation
SBM	BB To PSTN Ratio	Customer Retention	Customer Winback
RGM	Customer Segmentation	Collection Ratio	Capacity

Each feature is defined in-terms of certain ratio of two KPIs as mentioned in Table 2.

Table 2: Category Wise Features Definition

Feature	Category	Metric	Definition
COP	Co-BM	On-time Provisioning	On-Time Sales / Total Sales
CDN	Co-BM	Disconnection	Disconnection / Total Subs
CFR	Co-BM	Faults Registered/100/Month	Faults Registered / Total Subs
CRP	Co-BM	Repeat Faults	Faults Repeated / Total Faults
CRT	Co-BM	Repeat Telephones	Faults Repeated Telephone / Total Faults

COR	Co-BM	On Time Rectification	On-Time Rectification/Total Rectification
CDI	Co-BM	Customer Denied Installation	Denied Installation/Total Sales
CDR	Co-BM	Customer Denied Rectification	Denied Rectification/Total Dialed
CAN	Co-BM	Net Add Ratio	Total Sales / Disconnection
SBR	SBM	BB To PSTN Ratio	Total Subs BB / Total Subs PSTN
SCR	SBM	Customer Retention	Retained / Total Suspended
SCW	SBM	Customer Winback	Winback / Disconnection
RCS	RGM	Customer Segmentation	Segment / Total Subs
RCP	RGM	Capacity	Total Subs / Capacity
RCR	RGM	Collection Ratio	Recovered within due date / Total Recievable

3.1 Features of Co-Business Manager

Table 3 covers nine features of Co-Business Manager.

Table 3: Co-BM Features

COP	Co-BM	On-Time Provisioning
COR	Co-BM	On-Time Rectification
CFR	Co-BM	Faults Registered
CRP	Co-BM	Repeat Faults
CRT	Co-BM	Repeat Telephone
CAN	Co-BM	Net Adds
CDN	Co-BM	Disconnection
CDR	Co-BM	Customer Denied Rectification
CDI	Co-BM	Customer Denied Installation

3.1.1 On-Time Provisioning (COP)

The co-business managers are required to facilitate the customers by providing the service (new connection) with in the permissible timeline e.g. the landline connection within 3 days and broadband within 5 days.

COP feature is a ratio of total orders installed on-time to total orders installed.

*COP = [(Total On Time provisioned) / (Total Installed) * 100]* (4)

To calculate score as per business definition:

*Proportion = [(COP – Low Limit) / (Upper Limit – Low Limit) *Weightage]* (5)

3.1.2 On-Time Rectification (COR)

This KPI is to control the delays in faults rectification. Each complaint must be resolved within 24 hours.

COR feature is a ratio of total faults rectified within 24 hours to total faults registered.

*COR = [(Total On Time Rectified) / (Total Faults Registered) *100]* (6)

To calculate score as per business definition:

*Proportion = [(COR – Low Limit) / (Upper Limit – Low Limit) *Weightage]* (7)

3.1.3 Faults Registered (CFR)

CFR is the ratio between total faults registered and total number of active subscribers and its KPI (key performance indicator) is within 3%.

CFR feature is a ratio of total faults registered to total subscriber base.

*CFR = [(Total Faults Registered) / (Total Base) *100]* (8)

To calculate score as per business definition:

Proportion =

*[Weightage – ((CFR – Low Limit) / (Upper Limit – Low Limit) * Weightage)]* (9)

3.1.4 Repeat Faults (CRP)

The number of time any type of fault re-appear against each subscriber in last 15 days and the permissible repeat fault range is between 10% to 20%.

CRP feature is a ratio of total faults repeated to total faults registered.

*CRP = [(Total Faults Repeated) / (Total faults registered) *100]* (10)

To calculate score as per business definition:

Proportion =

*[Weightage – ((CRP – Low Limit) / (Upper Limit – Low Limit) * Weightage)]* (11)

3.1.5 Repeat Telephones (CRT)

The repeat telephone feature refers to the total number of unique subscribers (subscription wise) who are affected in services or have lounged the repeated complaints.

CRT feature is a ratio of total repeated telephone or customer to total faults registered.

*CRT = [(Total Repeated Telephone) / (Total faults registered) *100]* (12)

To calculate score of CRT as per business definition:

Proportion =

*[Weightage – ((CRT – Low Limit) / (Upper Limit – Low Limit) * Weightage)]* (13)

3.1.6 Net Additions (CAN)

The gross, net and churn targets are defined on yearly and monthly basis to find the BM's progress. These KPIs will be evaluated to compare the business manager's team progress or BM's progress.

CAN feature is a ratio of total repeated telephone or customer to total faults registered.

$$CAN = [(Total\ Gross\ Sales) / (Total\ Disconnections) *100] \quad (14)$$

To calculate score of CAN as per business definition:

$$Proportion = [(CAN - Low\ Limit) / (Upper\ Limit - Low\ Limit) *Weightage] \quad (15)$$

3.1.7 Disconnection (CDN)

If the rate of churn/disconnection of the subscribers is in control (as per KPI), the quality of service is ensured.

CDN feature is a ratio of total subscribers disconnected to total active base.

$$CDN = [(Total\ Subs\ Disconnected) / (Total\ Base) *100] \quad (16)$$

To calculate score of CDN as per business definition:

$$Proportion =$$

$$[Weightage - ((CDN - Low\ Limit) / (Upper\ Limit - Low\ Limit) * Weightage)] \quad (17)$$

3.1.8 Customer Denied Rectification (CDR)

Whenever a customer complaint is verified / rectified, the OBCC (Out Bound Call Center) calls customer to verify as if the complaint has really been fixed.

CDR feature is a ratio of total faults denied to total faults registered.

$$CDR = [(Total\ Faults\ Denied) / (Total\ Faults\ Registered) *100] \quad (18)$$

To calculate score of CDR as per business definition:

$$Proportion =$$

$$[Weightage - ((CDR - Low\ Limit) / (Upper\ Limit - Low\ Limit) * Weightage)] \quad (19)$$

3.1.9 Customer Denied Installation (CDI)

The OBCC solicits the subscriber's feedback on new connections to determine as if it's not the fake installation request? A business manager with 0 customer denied installation would get the maximum weight.

CDI feature is a ratio of total installation denied to total connections installed.

$$CDI = [(\text{Total denied Install}) / (\text{Total Installed}) * 100] \tag{20}$$

To calculate score of CDI as per business definition:

$$\text{Proportion} = [\text{Weightage} - ((CDI - \text{Low Limit}) / (\text{Upper Limit} - \text{Low Limit}) * \text{Weightage})] \tag{21}$$

3.2 Features of Senior Business Manager

Table 4 covers three features of Co-Business Manager.

Table 4: SBM Features

SBR	SBM	BB To PSTN Ratio
SCR	SBM	Customer Retention
SCW	SBM	Customer Winback

3.2.1 Broadband to PSTN Ratio (SBR)

The broadband to PSTN ratio is determined to find the total number of lines having broadband service. If ratio is more than 4%, the team is more productive.

SBR feature is a ratio of total broadband subscribers to total PSTN subscribers.

$$SBR = [(\text{Total DSL Subs}) / (\text{Total PSTN Subs}) * 100] \tag{22}$$

To calculate score of SBR as per business definition:

$$\text{Proportion} = [(SBR - \text{Low Limit}) / (\text{Upper Limit} - \text{Low Limit}) * \text{Weightage}] \tag{23}$$

3.2.2 Customer Retention (SCR)

TOS (temporary out of service), if a customer has not paid the bill the service is suspended temporarily. On payment the TOS connections are restored. The restoration ratio has direct impact on revenue.

SCR feature is a ratio of total restored subscribers to total TOS subscribers.

$$SCR = [(\text{Total Restored Subs}) / (\text{Total TOS Subs}) * 100] \tag{24}$$

To calculate score of SCR as per business definition:

$$\text{Proportion} = [(SCR - \text{Low Limit}) / (\text{Upper Limit} - \text{Low Limit}) * \text{Weightage}] \tag{25}$$

3.2.3 Customer Winback (SCW)

If a customer has not paid the bill or have left the service deliberately on request; the winback means to bring him/her back after churn.

SCW feature is a ratio of total restored subscribers to total TOS subscribers.
SCW = [(Total Winback Subs) / (Total Disconnections) *100] (26)

To calculate score of SCW as per business definition:
Proportion = [(SCW – Low Limit) / (Upper Limit – Low Limit) *Weightage] (27)

3.3 Features of Regional General Manager

Table 5 depicts three features of Regional General Manager (RGM).

Table 5: RGM Features

RCS	RGM	Customer Segmentation
RCP	RGM	Capacity
RCR	RGM	Collection Ratio

3.3.1 Customer Segmentation (RCS)

The subscribers are segmented into platinum plus, platinum, gold and silver based on

a) Average revenue of last six months
b) Level of service (data rate)
c) Customer longevity
d) In-time payments.

The total numbers of platinum plus, platinum, gold and silver customer of the respective regional general manager would be evaluated to predict a new manager as rising business manager. The RCS feature is a ratio of total valued customer (platinum, platinum plus and gold Subs) to total Active Subscribers.

RCS = [(Total Valued Subscribers) / (Total Active Subscribers) *100] (28)

To calculate score of SCW as per business definition:
Proportion = [(RCS – Low Limit) / (Upper Limit – Low Limit) *Weightage] (29)

3.3.2 Capacity (RCP)

It's RGM's KPI to manage the maximum capacity utilization and enhance the existing capacity of network elements. This is a ratio of total working connections and total capacity available on any exchange.

RCP feature is a ratio of total Active subscribers to total capacity of network element.

$$RCP = [(\text{Total Active Subs}) / (\text{Total Capacity}) * 100] \qquad (30)$$

To calculate score of SCW as per business definition:

$$\text{Proportion} = [(RCP - \text{Low Limit}) / (\text{Upper Limit} - \text{Low Limit}) * \text{Weightage}] \qquad (31)$$

3.3.3 Collection Ratio (RCR)

Its ratio of bills receivable to bills collected in-time. Every regional general manager needs to maintain the collection percentage as per KPI and it will be evaluated to determine the rising manager. It's very important KPI as if a subscriber does not pay the bill, the tax amount 19% is charged to the operator.

RCR feature is a ratio of total in-time collections to total receivable.

$$RCR = [(\text{Total in-time collection}) / (\text{Total receivables}) * 100] \qquad (32)$$

To calculate score of RCR as per business definition:

$$\text{Proportion} = [(RCR - \text{Low Limit}) / (\text{Upper Limit} - \text{Low Limit}) * \text{Weightage}] \qquad (33)$$

The table 6-point matrix represents the definition of each feature, the business target, weightage assigned to each feature and relationship type as if relation is direct or inverse. It further shows the formulae to calculate the proportional points.

Table 6: Point Matrix

Feature	Ratio	KPI (Target)	Weightage	Low (0 points)	High (Max Points)	Proportional Points
On-time Provisioning	On-Time Sales/Total Sales	PSTN: >=90% in 3 days BB: >=90% in 5 days	10	<=60%	>=90%	((V-Y)/(X-Y))*W
Disconnection	Disconnection/Total Subs	<=1%	10	>=4%	<=1%	W-((V-Y)/(X-Y)*W)
Faults Registered /100/Month	Faults Reg/ Total Subs	<=3.08 Faults	10	>=12	<=3.08	W-((V-Y)/(X-Y)*W)

Repeat Faults	Faults Rep/Total Faults	<=10%	10	>=30%	<=10%	W-((V-Y)/(X-Y)*W)
Repeat Telephones	Faults Rep Tel/Total Faults	<=10%	10	>=20%	<=10%	W-((V-Y)/(X-Y)*W)
On Time Rectification	On-Time Rectified / Total Rectified	>=95% in 24 hrs	10	<=50%	>=95%	((V-Y)/(X-Y))*W
Customer Denied Installation	Denied Install/Total Sales	<=0%	10	>=5%	<=0%	W-((V-Y)/(X-Y)*W)
Customer Denied Rectification	Denied Rectified / Total Dialed	<=10%	10	>=30%	<=10%	W-((V-Y)/(X-Y)*W)
BB To PSTN Ratio	Total Subs BB / Total Subs PSTN	>=70	10	<=30	>=70	((V-Y)/(X-Y))*W
Customer Retention	Restored/Total TOS	>=60%	10	<=40%	>=60%	((V-Y)/(X-Y))*W
Customer Winback	Winback/Disconnection	>=25%	10	<=15%	>=25%	((V-Y)/(X-Y))*W
Customer Segmentation	(Platinum + Platinum plus + Gold) / Total Subs	>=30%	10	<=20%	>=30%	((V-Y)/(X-Y))*W
Capacity	Total Subs / Capacity	>=65%	10	<=40%	>=65%	((V-Y)/(X-Y))*W
Net Add Ratio	Total Sales/Disconnection	>=70%	10	<=50%	>=70%	((V-Y)/(X-Y))*W
Collection Ratio	Recovered within due date/Total Recievable	>=40%	10	<=20%	>=40%	((V-Y)/(X-Y))*W

CHAPTER 4

Experiment

4. Experiment

1. The dataset has been engaged from a leading telecom company **PTCL** for 13 months (Sep 2014 to Sep 2015) to predict rising business manager.

2. The achievements/points of co-business manager and overall team is evaluated against KPIs based on progress and proportionate points. The variance of achievements is determined based on ratio of maximum and minimum points of last 3 months.

3. The discriminative learning framework is used to determine the rising business manager in dataset against each KPI.

4. Total features are fifteen wherein nine features relate to co-business manager.

4.1 Dataset

The dataset will be acquired from PTCL for a maximum period of last 1 year to predict the new business manager as **rising** based on the performance of co-business managers. The profile and working jurisdiction will be as per latest position. As per working hierarchy of the business manager, the Business Manager (BM) reports to Senior Business Manager (SBM) and SBM is reporting to Regional General Manager (RGM). The total RGMs are 18 (one for each region) and total BMs falls around 170. However, the total SBMs are at most 70. This research work will carry the dataset of two major telecom products i.e. PSTN (landline) and Broadband.

The dataset for the period Sep 2014 to Sep 2015 have been acquired from the database to predict rising business manager. Each feature is a ratio of two correlated key performance indicators. The total numbers of feature are fifteen and sixteenth feature is class label. The class label is nominal independent variable that marks business manager as rising or not rising based on revenue.

The total instances or business managers are 175 and based on revenue top 50 business managers are marked as rising and rest of 125 business managers are labeled as not rising. The six months aggregated data for the period Sep 2014 to Feb 2015 has been sliced randomly into 10 sample data sets (10, 20, 30...100) of equal rising and not rising business managers.

4.2 Performance Evaluation

The random data sets (10, 20, 30...100) having equal instances of binary class label are explored in Weka 3.7 (Waikato Environment for Knowledge Analysis) with Neural Net, SVM, Bayes Net and Naïve Bayes classifiers and 5-fold cross validation is applied on data sets. IBM SPSS Modeler is also used for results evaluation through expert modeling technique.

The 5-fold cross validation divides the sample data set into five equal parts. The four parts are used for training while one part is used for testing purpose. This process is repeated five times and each time different five sample parts are used for testing. Then the average result rate is determined. The accuracy, precision, recall and f-measure are determined after applying five-fold cross validation on each sample data set.

4.2.1 Accuracy

The accuracy of a system is the degree of closeness of dimensions of a number to that numbers real value [22]. In nutshell, the accuracy is how close the measured values are to the actual values. The accuracy is a part of accurate results in a population for both true positive and true negative. It is the test parameters [23].

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{34}$$

tp stands for true positives	tn stands for true negatives
fp stands for false positives	fn stands for false negatives

4.2.2 Precision

The precision of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results. In other words, the precision is how close the measured values are to each other.

The precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives) [23].

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{35}$$

4.2.3 Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a dataset. It is usually called sensitivity, and corresponds to the true positive rate. It is defined by the formula [23]:

$$Recall = sensitivity = \frac{true\ positive}{(true\ positive + false\ negative)} \tag{36}$$

True positive + false negative is total no. of analysis instances of the measured class.

4.2.4 F-Measures

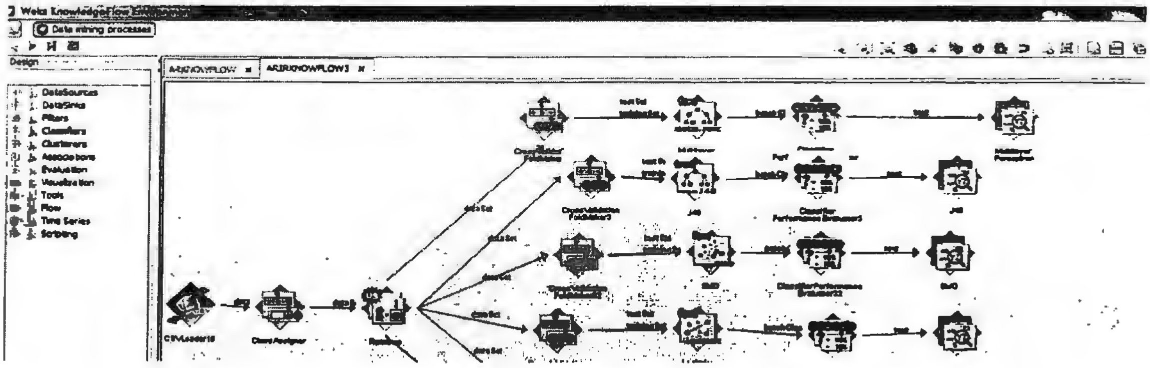
A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score [23]:

$$F = 2.\frac{\text{precision. recall}}{\text{precision} + \text{recall}} \tag{37}$$

This is also known as the F1 measure, because recall and precision are evenly weighted.

4.3 Implementation of Classification Models in WEKA

The Waikato Environment for Knowledge Analysis (WEKA) is renowned as a most widely used tool for research purpose in data mining and has grasped broad recognition in the academic circles and industries [24]. Weka is a set of machine learning and data mining algorithms. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka controls tools for classification, regression, data pre-processing, clustering, association rules, and visualization. Weka is open source and freely available software issued under the GNU (General Public License). Figure 4-1 depicts a snapshot of Weka workflow which is helpful tool to retrieve results from all classifiers simultaneously and to export results in text file for further processing.



4-1: Weka Workflow

4.3.1 Results and Discussion

This section provides the detailed results of classification modeling techniques that we have been used for finding rising business manager. We have also performed category wise and features wise results discussion.

Table 7: Implementation Filters

Action	Data Set Size	Features	Classifiers
Data Set Size Selection	50	All	SVM, NN, NB, BN
Feature Sub Set Selection	50	RCR, CDN, RCP,CDR,COR,CFR	SVM, NN, NB, BN

Individual Feature Analysis	50	RCR,CDN, RCP,CDR,COR,CFR	SVM, NN, NB, BN
Category Wise Analysis	50	All	SVM, NN, NB, BN
Model Wise Combined Feature Analysis	10 to 100	All	SVM, NN, NB, BN
Baseline Analysis	50	All	SVM, NN, NB, BN

4.3.2 Data Set Size Selection

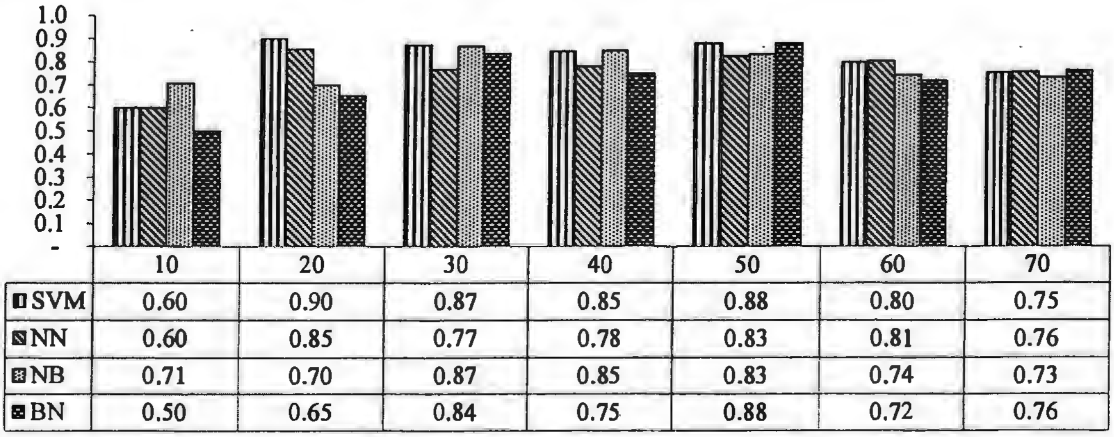
Each classifier is executed on randomly selected sample data sets of 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 for AR and ARIR to determine the data set size having the best average results. The selected data set is then used in all later evaluations.

Scenario 1: Data Set Size Selection Based on Average Revenue (AR)

In this section we analyzed precision, recall and f-measures of each AR data set to find the data set size having best average results against all classifiers. We also ensured that the results afterwards are actually declining.

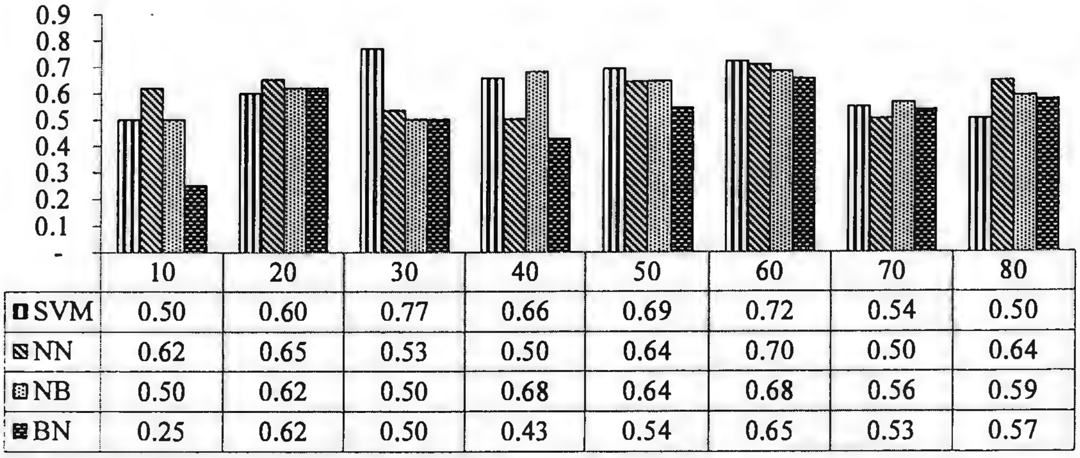
Result: Precision Analysis of Data Set Size Selection

Figure 4-2 show the precision results of AR data sets by using four different classification models (SVM, NN, NB, BN). The comparison of results against different data sets shows that the best average precision against all models are on data set 50 i.e. 85%. However, on data set size afterwards like 60 and 70 the precision gets declined to 77% and 75% respectively. This has also been observed that SVM and BN classifiers outperform equally with precision of 88% on data set size 50 that has yielded best results.



	10	20	30	40	50	60	70
▥SVM	0.60	0.90	0.87	0.85	0.88	0.80	0.75
▧NN	0.60	0.85	0.77	0.78	0.83	0.81	0.76
▦NB	0.71	0.70	0.87	0.85	0.83	0.74	0.73
■BN	0.50	0.65	0.84	0.75	0.88	0.72	0.76

4-2: Precision Analysis of Data Set Selection

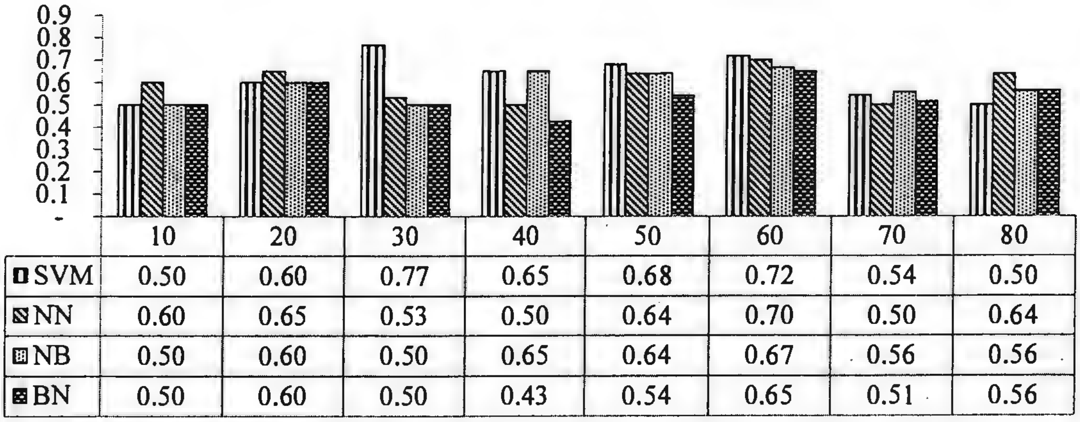


	10	20	30	40	50	60	70	80
□ SVM	0.50	0.60	0.77	0.66	0.69	0.72	0.54	0.50
▧ NN	0.62	0.65	0.53	0.50	0.64	0.70	0.50	0.64
▤ NB	0.50	0.62	0.50	0.68	0.64	0.68	0.56	0.59
▦ BN	0.25	0.62	0.50	0.43	0.54	0.65	0.53	0.57

4-5: Precision Analysis of Data Set Selection Using ARIR

Result: Recall Analysis of Data Set Size Selection

Likewise, Figure 4-6 shows the recall analysis of ARIR data sets by using four different classification models. The comparison of results against different data sets shows that the best average recall against all models are on data set 60 i.e. 68%. However, on data set size afterwards like 70 the recall gets declined to 53%. The SVM classifier outperforms with precision of 72% on selected data set size i.e. 60.

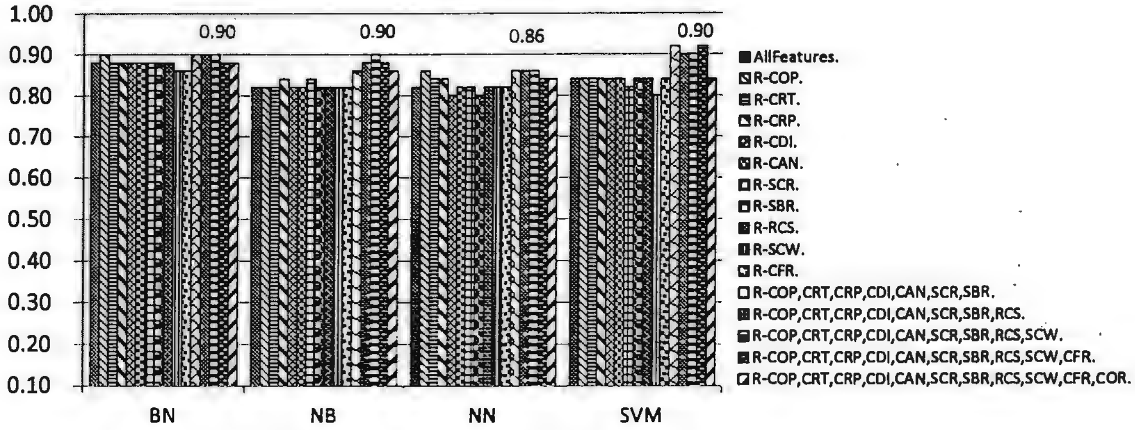


	10	20	30	40	50	60	70	80
□ SVM	0.50	0.60	0.77	0.65	0.68	0.72	0.54	0.50
▧ NN	0.60	0.65	0.53	0.50	0.64	0.70	0.50	0.64
▤ NB	0.50	0.60	0.50	0.65	0.64	0.67	0.56	0.56
▦ BN	0.50	0.60	0.50	0.43	0.54	0.65	0.51	0.56

4-6: Recall Analysis of Data Set Selection Using ARIR

Result: F-Measure Analysis of Data Set Size Selection

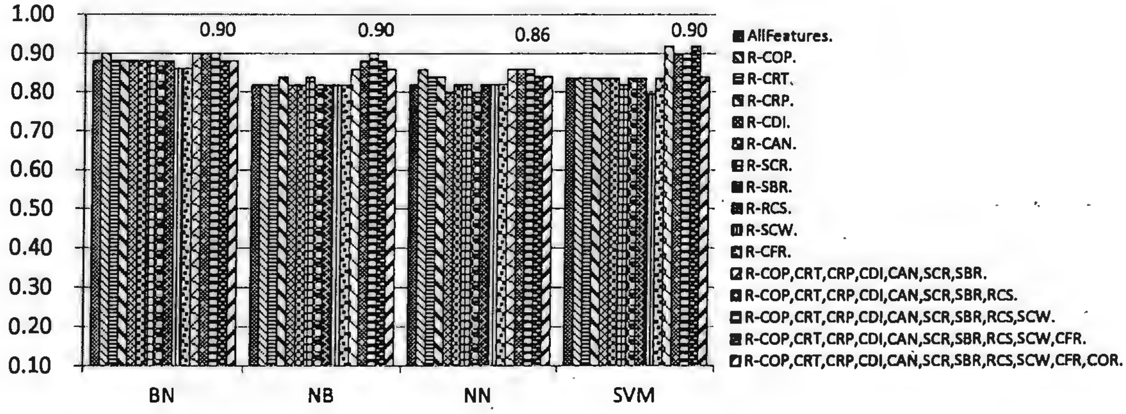
Similarly, Figure 4-7 depicts the f-measure analysis of ARIR data sets by using four different classification models. The comparison of results against different data sets shows that the best average f-measure against all models are on data set 60 i.e. 68%. However, on data set size afterwards like 70 the recall gets declined to 51%. The SVM classifier outperforms with precision of 72% on selected data set size i.e. 60.



4-9: Recall Analysis of Feature Subset Selection Using AR

Result: F-Measure Analysis of Features Evaluation

The F-Measure results of feature subset selection are compared with actual results as shown in figure 4-10. The actual results are the average results retrieved from all features on 50 instances data set. The result shows that if COP, CRT, CRP, CDI, CAN, SCR, SBR, RCS and SCW features are removed from data set, the results will get better with f-measure weighted average of 89% but if next least significant feature as per output of feature evaluation algorithms is dropped like CFR or COR; the outcome will decline. The result also indicates that RCR cannot be used as only one feature for rising star prediction as it yields results upto 76%.



4-10: F-Measure Analysis of Feature Subset Selection Using AR

Scenario 2: Feature Subset Selection Based on Average Relative Increase in Revenue (ARIR)

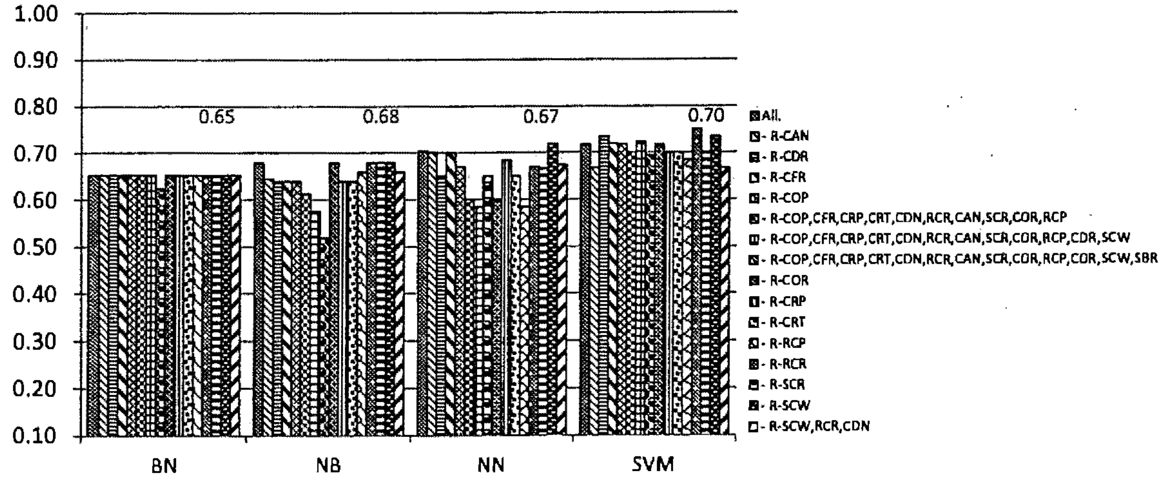
The result in matrix below shows that SCW feature is less correlated and by omitting this feature the results will be improved.

Table 9: Feature Sub Set Selection (ARIR)

Info Gain on Full Training Set			Gain Ratio on Full Training Set			Chi Square on Full Training Set		
Ranking Score	**Feature Sequence**	**Feature Abv.**	**Ranking Score**	**Feature Sequence**	**Feature Abv.**	**Ranking Score**	**Feature Sequence**	**Feature Abv.**
0.23	14	CDI	0.294	3	SBR	17.33	14	CDI
0.196	6	RCS	0.256	14	CDI	15.429	6	RCS
0.191	3	SBR	0.2	6	RCS	12	3	SBR
0	5	SCW	0	5	SCW	0	5	SCW
0	15	CDR	0	15	CDR	0	15	CDR
0	7	RCP	0	7	RCP	0	7	RCP
0	2	COR	0	2	COR	0	2	COR
0	4	SCR	0	4	SCR	0	4	SCR
0	8	CAN	0	8	CAN	0	8	CAN
0	9	RCR	0	9	RCR	0	9	RCR
0	10	CDN	0	10	CDN	0	10	CDN
0	13	CRT	0	13	CRT	0	13	CRT
0	12	CRP	0	12	CRP	0	12	CRP
0	11	CFR	0	11	CFR	0	11	CFR
0	1	COP	0	1	COP	0	1	COP

Result: Precision Analysis of Feature Subset Selection

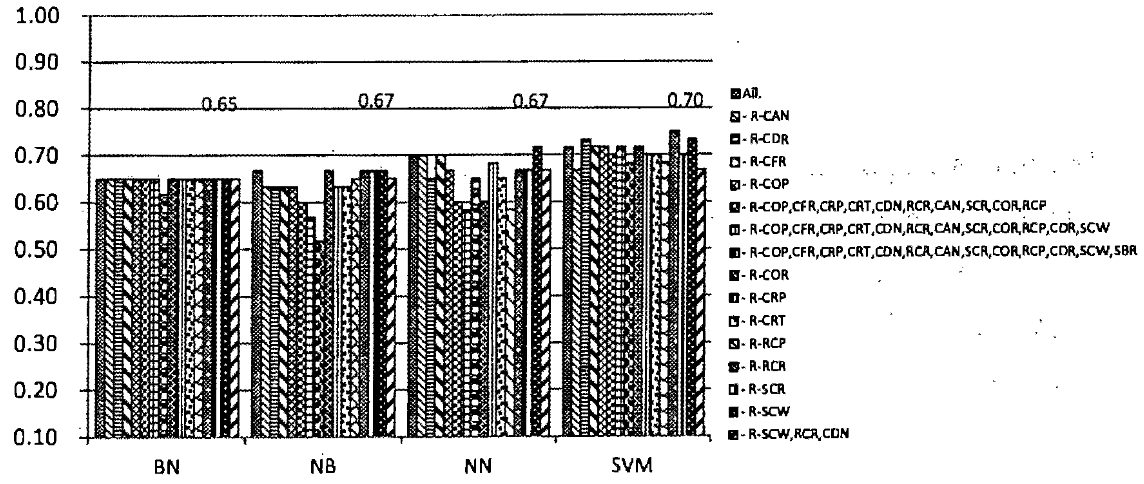
The precision results of feature subset selection are compared with actual results as shown in figure 4-11. The actual results are the average results of all features retrieved on 60 instances data set. The result shows that if SCW feature is removed from data set, the results will get better with precision weighted average of 70% but if next least significant feature as per output of feature evaluation algorithms is dropped; the outcome will decline. The result also indicates that CDI cannot be used as only one feature for rising star prediction as it yields results upto 66%.



4-11: Precision Analysis of Feature Subset Selection Using ARIR

Result: Recall Analysis of Feature's Evaluation

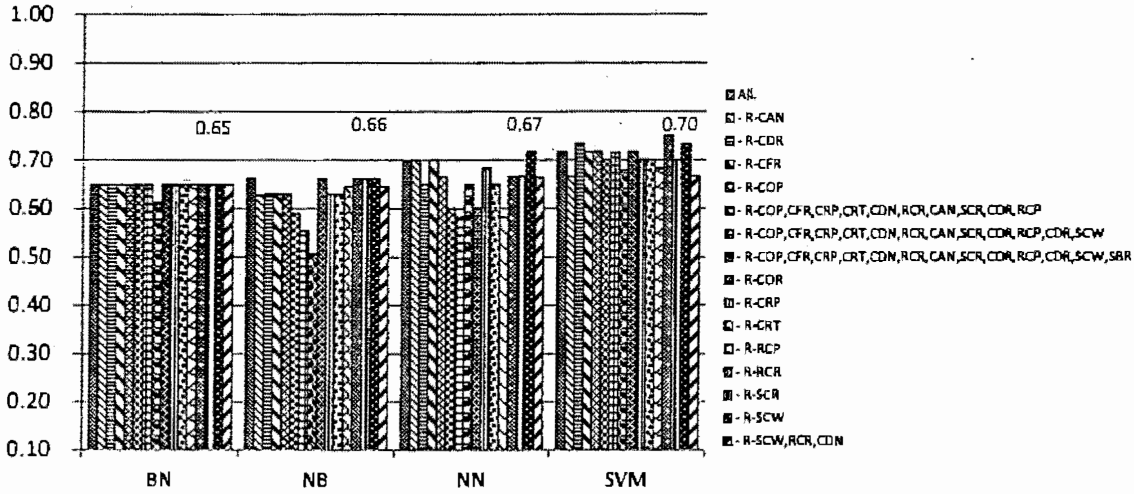
The recall results of feature subset selection are compared with actual results as produced in figure 4-12. The actual results are the average results of all features retrieved on 60 instances data set. The result shows that if SCW feature is removed from data set, the results will get better with precision weighted average of 69% but if next least significant feature as per output of feature evaluation algorithms is dropped; the outcome will decline. The result also indicates that CDI cannot be used as only one feature for rising star prediction as it yields results upto 64%.



4-12: Recall Analysis of Feature Subset Selection Using ARIR

Result: F-Measure Analysis of Features Evaluation

The F-Measure results of feature subset selection are compared with actual results as shown in figure 4-13. The actual results are the average results of all features retrieved on 60 instances data set. The result shows that if SCW feature is removed from data set, the results will get better with precision weighted average of 69% but if next least significant feature as per output of feature evaluation algorithms is dropped; the outcome will decline. The result also indicates that CDI cannot be used as only one feature for rising star prediction as it yields results upto 63%.



4-13: Precision Analysis of Feature Subset Selection Using ARIR

4.3.4 Individual Features Analysis Using Classification Models

We have calculated precision, recall and f-measures of all features by using classification models such as BN, NB, NN and SVM on sample data set of size 50 for period Sep 2014 to Feb 2015. In the first type of dataset, the business managers are selected for performance analysis on the basis of weighted average of revenue (class label) and in the second type of dataset, the business managers are selected for performance analysis on the basis of Average Relative Increase in revenue (ARIR). The ARIR is derived similarly as [7].

$$\text{ARIR} = \max_{i \in T} \text{Change}_i * P_L * \sum_{i \in T} \text{Change}_i \qquad (38)$$

Where PL is last standing revenue i.e. of months Sep 2015, T is total no of months and Change i is the increase in revenue for current year i. The Change i is formulated as:

Change i = (Pi - P (i-1)) / Pi (39)

The NN classifier outperforms when performance of individual feature is evaluated because it applies weight on sub elements of each feature. Whereas, the SVM classifier at least require two features and it builds vector support among all features. Since SVM classifier builds a vector so the features that are not linearly separable are best candidate for SVM. The NB and BN classifiers mainly relies on Info Gain and Gain Ratio of each feature. The Info Gain and Gain Ratio are feature selection techniques so the feature with low correlation are not considered. In contrast, PCA (principal component analysis) is feature reduction technique that takes into account the combination of features.

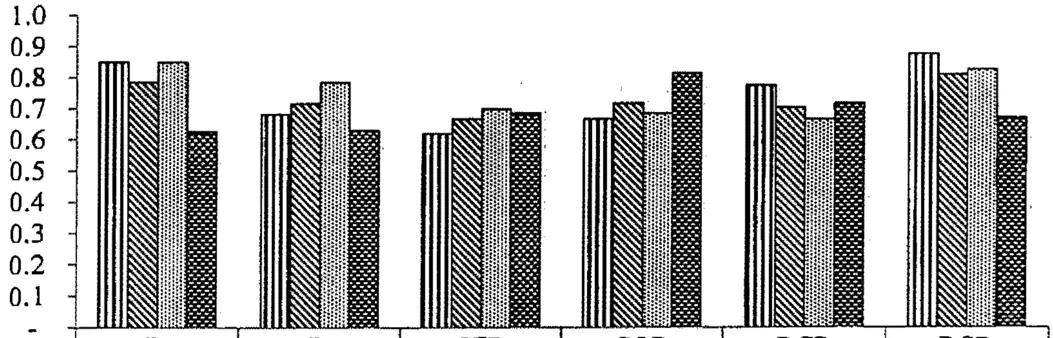
Scenario 1: Individual Features Analysis Based on Average Revenue (AR)

In this section we analyze precision, recall and f-measure of individual feature of each business manager by using sample dataset of size 50 which is the best case in AR data set. The feature set is of 1st six months from Sep 2014 to Feb 2015 whereas the class label is based on average revenue of all months from Sep 2014 to Sep 2015. As mentioned in section 4.2, the 5-fold cross validation method is selected for classification of classifier's training and validation.

Result: Precision Analysis of Features

Figure 4-14 shows the precision analysis result of feature by using classification modeling techniques. In figure 1 we observed that:

1. The feature collection ratio (RCR) has produced the highest average results of accuracy 79% using all six classifiers. The BN classifier has produced the best accuracy of 88% against RCR feature for prediction of rising business manager.
2. The disconnection (CDN) feature has produced the second highest average result of accuracy 78% using NN, BN, NB and SVM algorithms.
3. Among all features, the most correlated features are RCR and CDN that carries average impact of 79% and 78% respectively.
4. Against all features, the Neural Net classifier has produced the maximum average results of 75%.
5. On individual feature level, the generative nature classifiers produce better average accuracy results of 74% as compared to discriminative classifiers having 72%

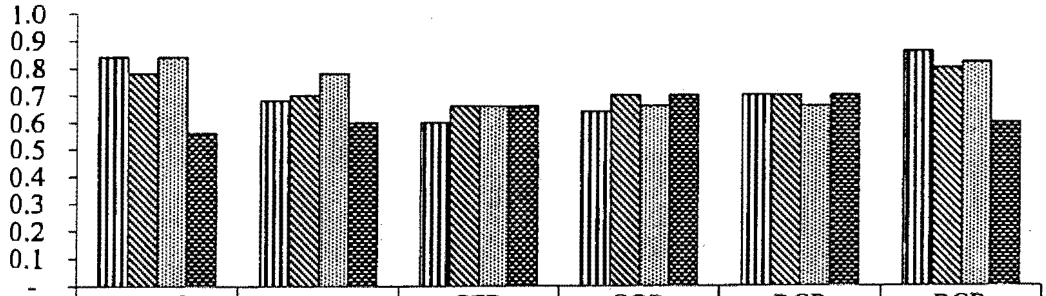


	CDN	CDR	CFR	COR	RCP	RCR
◻BN	0.85	0.68	0.62	0.67	0.77	0.88
▧NB	0.78	0.72	0.67	0.72	0.70	0.81
▤NN	0.85	0.78	0.70	0.68	0.67	0.83
▩SVM	0.63	0.63	0.68	0.81	0.72	0.67

4-14: Features Precision Analysis Using AR

Result: Recall Analysis of Features

Similarly figure 4-15 show the recall result of features using classification models. In figure 2, it's depicted that feature collection ratio has produced highest average accuracy of 77% and with BN algorithm the same feature shows 86% accuracy which is the highest accuracy. However, the NN classifier has produced best average recall of 74%. The BN and NB classifiers carry next best and equal accuracy of 72%.



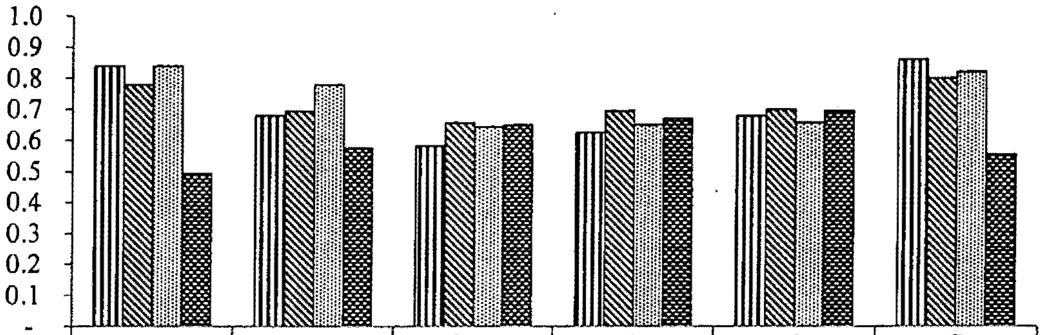
	CDN	CDR	CFR	COR	RCP	RCR
◻BN	0.84	0.68	0.60	0.64	0.70	0.86
▧NB	0.78	0.70	0.66	0.70	0.70	0.80
▤NN	0.84	0.78	0.66	0.66	0.66	0.82
▩SVM	0.56	0.60	0.66	0.70	0.70	0.60

4-15: Features Recall Analysis Using AR

Result: F-Measure Analysis of Features

Figure 4-16 show the F1 results of features by using classification modeling techniques. In figure 3 the findings are produced as:

1. The feature collection ratio (RCR) has produced the highest average results of accuracy 76% using all four classifiers. The BN classifier has produced the best accuracy of 86% against RCR feature for prediction of rising business manager.
2. The disconnection (CDN) feature has produced the second highest average result of accuracy 74% using BN, NB, NN and SVM classifiers.
3. Among all features, the most correlated features are RCR and CDN that have average impact of 76% and 74% respectively.
4. Against all features, the NN classifier has produced the maximum average results of 73%.
5. The generative nature classifiers produce better average accuracy results of 72% as compared to discriminative classifiers 67%.



	CDN	CDR	CFR	COR	RCP	RCR
□BN	0.84	0.68	0.58	0.63	0.68	0.86
▧NB	0.78	0.69	0.66	0.69	0.70	0.80
▤NN	0.84	0.78	0.64	0.65	0.66	0.82
▩SVM	0.49	0.58	0.65	0.67	0.69	0.55

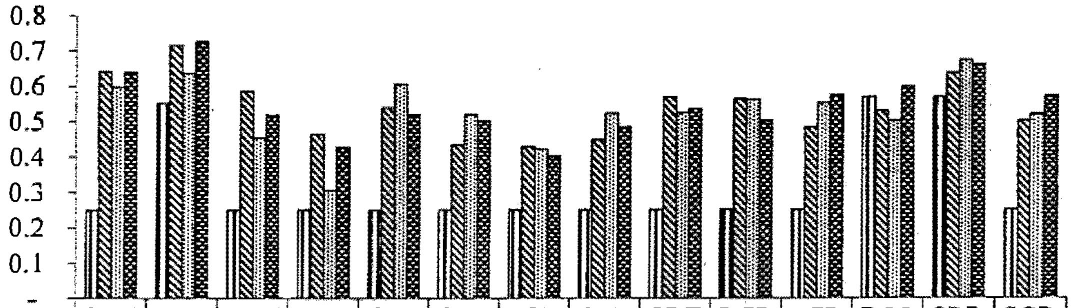
4-16: Features F-Measure Analysis Using AR

Scenario 2: Individual Features Analysis Based on Average Relative Increase in Revenue (ARIR)

In this section we analyze precision, recall and f-measure of features by using 2nd dataset Average Relative Increase (ARI). The feature set is of 1st six months from Sep 2014 to Feb 2015 whereas the class label is based on average relative increase of months from Mar 2015 to Sep 2015.

Result: Precision Analysis of Features

Figure 4-17 show the precision result of features by using 2^{nd} dataset (ARIR) Average Relative Increase in revenue. In figure 4, we observed that the CDI feature depicts highest average result of accuracy 66% and SVM model contributes the most as 73% in the same feature. The best classifier against all features in ARIR is also SVM with accuracy 55%.

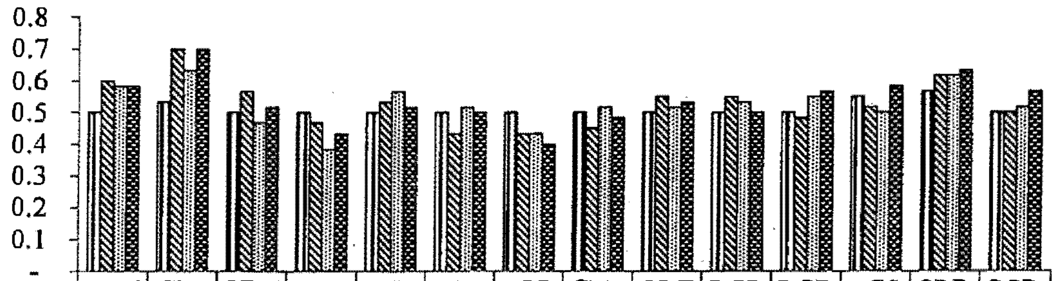


	CAN	CDI	CDN	CDR	CFR	COP	COR	CRP	CRT	RCP	RCR	RCS	SBR	SCR
□ BN	0.25	0.55	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.57	0.57	0.25
▨ NB	0.64	0.72	0.59	0.46	0.54	0.43	0.43	0.45	0.57	0.56	0.48	0.53	0.64	0.50
▤ NN	0.60	0.64	0.45	0.31	0.60	0.52	0.42	0.52	0.52	0.56	0.55	0.50	0.67	0.52
▣ SVM	0.64	0.73	0.52	0.43	0.52	0.50	0.40	0.48	0.53	0.50	0.57	0.60	0.66	0.57

4-17: Features Precision Analysis Using ARIR

Result: Recall Analysis of Features

Similarly figure 4-18 show the recall result of features using classification models. In figure 5, we observed that the CDI feature depicts highest average result of accuracy 64% whereas the SBR feature is next to it with 61% accuracy. The CDI feature has produced the better results using SVM and NB classifiers i.e. 70%.

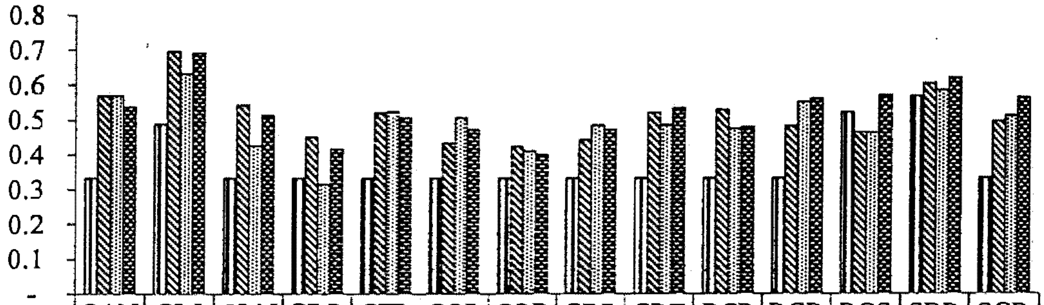


	CAN	CDI	CDN	CDR	CFR	COP	COR	CRP	CRT	RCP	RCR	RCS	SBR	SCR
□ BN	0.50	0.53	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.55	0.57	0.50
▨ NB	0.60	0.70	0.57	0.47	0.53	0.43	0.43	0.45	0.55	0.55	0.48	0.52	0.62	0.50
▤ NN	0.58	0.63	0.47	0.38	0.57	0.52	0.43	0.52	0.52	0.53	0.55	0.50	0.62	0.52
▣ SVM	0.58	0.70	0.52	0.43	0.52	0.50	0.40	0.48	0.53	0.50	0.57	0.58	0.63	0.57

4-18: Features Recall Analysis Using ARIR

Result: F-Measure Analysis of Features

Figure 4-19 show the F-Measure result of features by using classification models on Average Relative Increase in revenue (ARIR). It shows that the CDI feature depicts highest average result of accuracy 63%. The SBR feature has produced 2nd best average results of 59%. The SVM classifier outperforms with average accuracy of 52%. The NN model is next to SVM with average accuracy of 51%.



	CAN	CDI	CDN	CDR	CFR	COP	COR	CRP	CRT	RCP	RCR	RCS	SBR	SCR
□ BN	0.33	0.49	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.52	0.57	0.33
▧ NB	0.57	0.70	0.54	0.45	0.52	0.43	0.42	0.44	0.52	0.53	0.48	0.46	0.60	0.50
▦ NN	0.57	0.63	0.43	0.32	0.52	0.51	0.41	0.48	0.48	0.48	0.55	0.46	0.58	0.51
▩ SVM	0.54	0.69	0.51	0.42	0.51	0.47	0.40	0.47	0.53	0.48	0.56	0.57	0.62	0.56

4-19: Features F-Measure Analysis Using ARIR

4.3.5 Category Wise Analysis

In category wise analysis, the total feature set is segregated categorically into three i.e. co-BM, SBM and RGM. All four classifiers NB, BN, NN and SVM are applied on AR and ARIR data set of 50 and 60 instances respectively.

Table 10: Features Categories (AR)

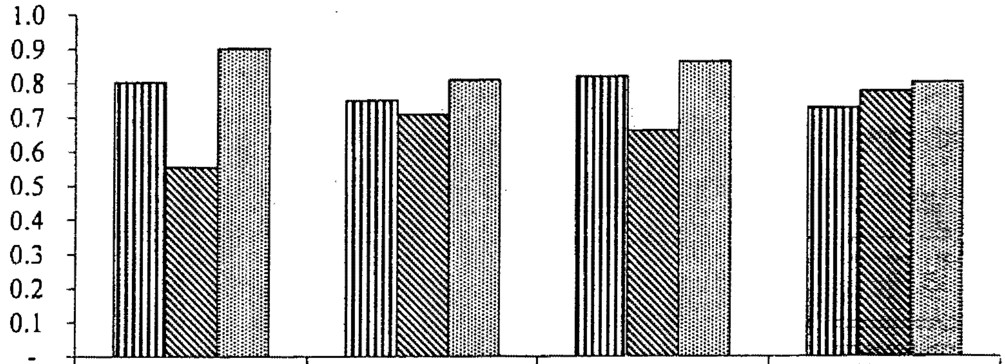
Category 1	**Co-BM**	On-Time Provisioning	On-Time Rectification	Faults Registered
		Repeat Faults	Repeat Telephone	Net Adds
		Disconnection	Cust. Denied Rectification	Cust Denied Installation
Category 2	**SBM**	BB To PSTN Ratio	Customer Retention	Customer Winback
Category 3	**RGM**	Customer Segmentation	Collection Ratio	Capacity

Scenario 1: Category wise Analysis Based on Average Revenue (AR)

In this section, we evaluated category wise precision, recall and f-measure of AR sample sets.

Result: Precision Analysis of Features Categories

The figure 4-20 shows the precision results of features categories by using classification models on the sample data set of 50 instances. The NN classifier outperforms with average accuracy of 78% and the SVM model carries the 2nd best average accuracy of 77%. Whereas, the RGM's feature set has produced the best average results of 84%.

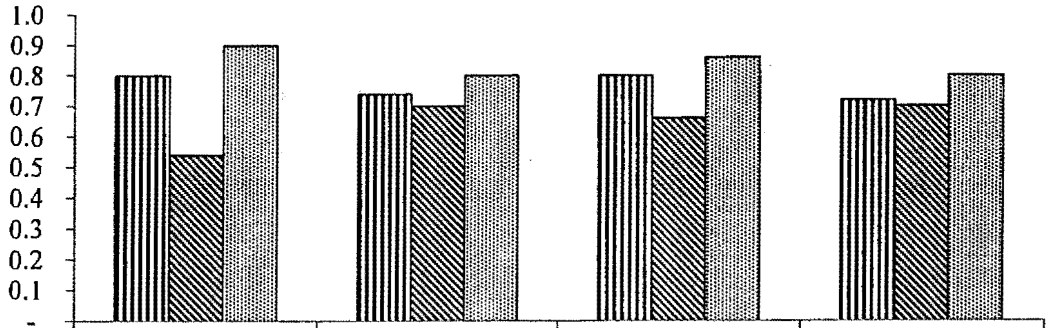


	BN	NB	NN	SVM
□ Co-BM	0.80	0.75	0.82	0.73
▧ SBM	0.56	0.71	0.66	0.77
▦ RGM	0.90	0.81	0.86	0.80

4-20: Category Wise Precision Analysis Using AR

Result: Recall Analysis of Features Categories

The figure 4-21 shows the recall results of features categories. The Neural Net classifier outperforms with average accuracy of 77% and the generative models (BN, NB) carries the 2nd best average accuracy of 75%. Whereas, the RGM's feature set has produced the best average results of 86%. The RGM features with BN classifier has maximum accuracy of 90%.

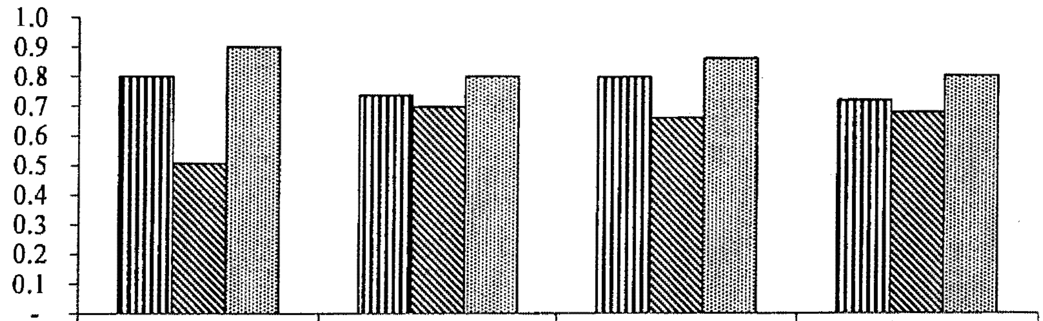


	BN	NB	NN	SVM
□ Co-BM	0.80	0.74	0.80	0.72
▧ SBM	0.54	0.70	0.66	0.70
▦ RGM	0.90	0.80	0.86	0.80

4-21: Category Wise Recall Analysis Using AR

Result: F-Measure Analysis of Features Categories

The figure 4-22 shows the f-measure results of features categories. The Neural Net classifier outperforms with average accuracy of 77% and the generative classifiers (BN, NB) carries the 2nd best average accuracy of 74%. Whereas, the RGM's feature set has produced the best average results of 84%. The RGM features with BN classifier carries best results of 90%.



	BN	NB	NN	SVM
□ Co-BM	0.80	0.74	0.80	0.72
▧ SBM	0.51	0.70	0.66	0.68
▦ RGM	0.90	0.80	0.86	0.80

4-22: Category Wise F-Measure Analysis Using AR

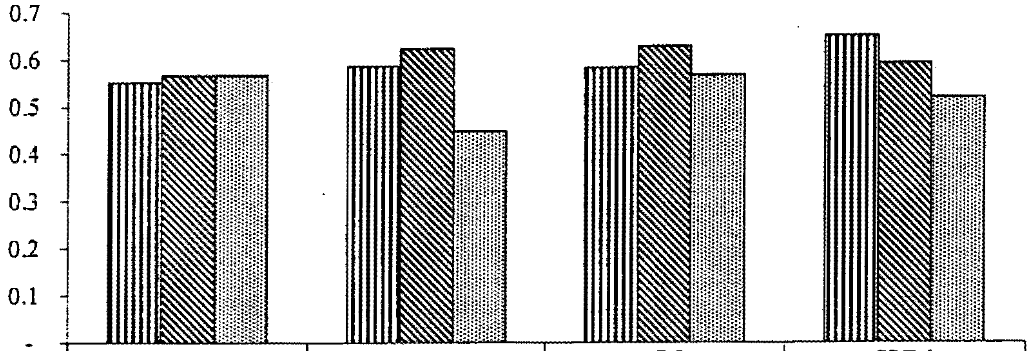
Scenario 2: Category wise Analysis Based on Average Relative Increase in Revenue (ARIR)

In this section, we weighed the category wise precision, recall and f-measure of ARIR sample set of size 60.

Result: Precision Analysis of Features Categories

The figure 4-23 shows the precision results of features categories by using classification models on the sample data set of 60 instances.

The SVM and NN classifier outperforms with same average accuracy of 59% and the BN model carries the 2nd best average accuracy of 56%. Whereas, the SBM's feature set has produced the best average results of 60%.



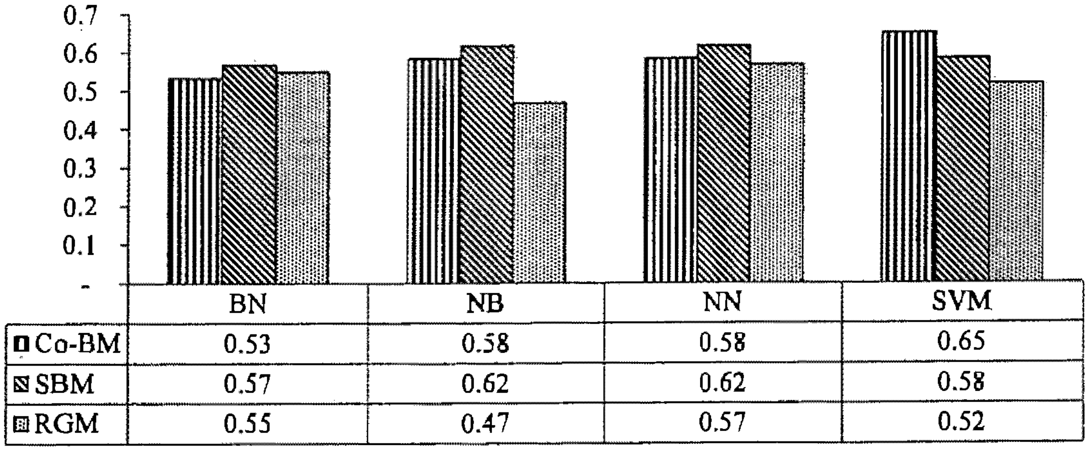
	BN	NB	NN	SVM
▥ Co-BM	0.55	0.59	0.58	0.65
▧ SBM	0.57	0.62	0.63	0.59
▦ RGM	0.57	0.45	0.57	0.52

4-23: Category Wise Precision Analysis Using ARIR

Result: Recall Analysis of Features Categories

The figure 4-24 shows the recall results of features categories by using classification models on the sample data set of 60 instances.

The NN classifier outperforms with average accuracy of 59% and the SVM model carries the 2nd best average accuracy of 58%. Whereas, the SBM's feature set has produced the best average results of 60%.



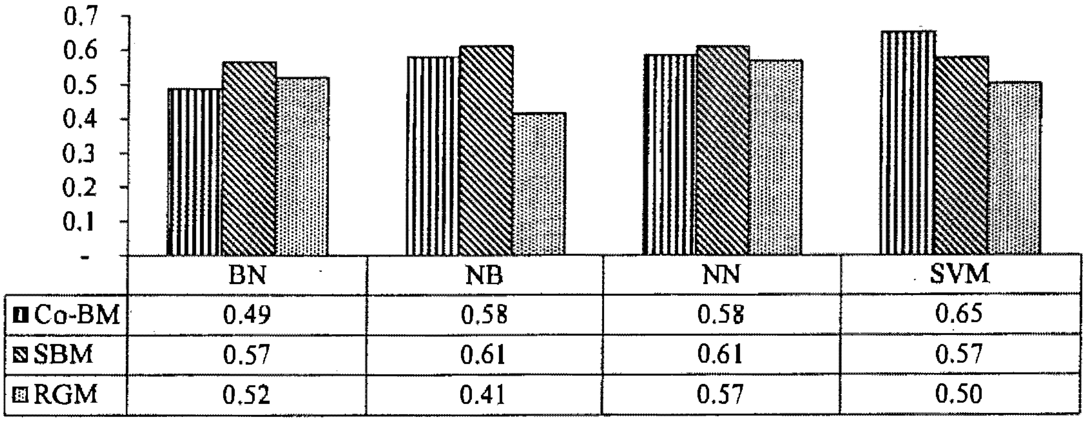
	BN	NB	NN	SVM
□ Co-BM	0.53	0.58	0.58	0.65
▧ SBM	0.57	0.62	0.62	0.58
▨ RGM	0.55	0.47	0.57	0.52

4-24: Category Wise Recall Analysis Using ARIR

Result: F-Measure Analysis of Features Categories

The figure 4-25 shows the f-measure results of features categories by using classification models on the sample data set of 60 instances.

The NN classifier outperforms with average accuracy of 59% and the SVM model carries the 2nd best average accuracy of 57%. Whereas, the SBM's feature set has produced the best average results of 59%.



	BN	NB	NN	SVM
□ Co-BM	0.49	0.58	0.58	0.65
▧ SBM	0.57	0.61	0.61	0.57
▨ RGM	0.52	0.41	0.57	0.50

4-25: Category Wise F-Measure Analysis Using ARIR

4.3.6 Model Wise Combined Features Analysis [All Features]

Each classifier is executed against all features on randomly selected sample data sets of 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 for AR and ARIR.

The SVM classifier builds a vector so the features that are not linearly separable are best candidate for SVM. The SVM classifier outperforms when number of features are more as in case of model wise combined feature analysis. The NB and BN classifiers mainly relies on Info Gain and Gain Ratio of each feature. The Info Gain and Gain Ratio are feature selection techniques so the feature with low correlation are not considered. In contrast, PCA (principal component analysis) is feature reduction technique that takes into account the combination of features.

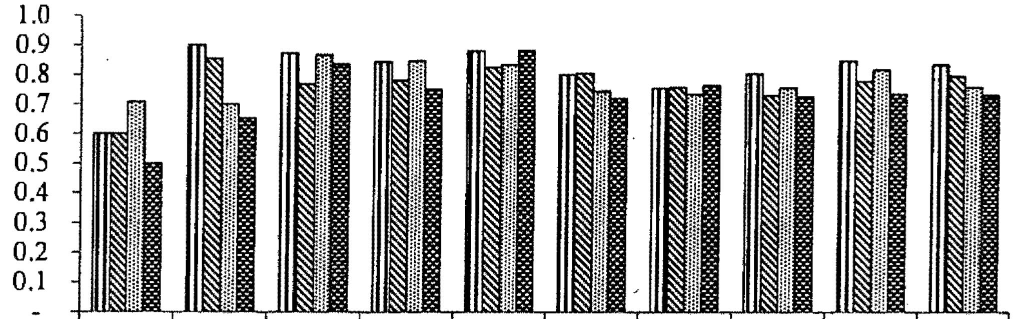
Scenario 1: Combined Features Analysis Based on Average Revenue (AR) – All Features

In this section we analyzed precision, recall and f-measures of features by using classification models on first type of datasets i.e. AR.

Result: Model Wise Precision Analysis of Features – All Features

Figure 4-26 show the precision results of all features by using four different classification models. The comparison of results between different classification models shows the performance of SVM model is better than other models. The SVM model offers the highest average accuracy result of 90%. The SVM model outperforms with average of 90% on data set of 20 instances. The BN classifier show 2nd best average precision of 88%.

The sample data set 10 carries worst average results of 60% and on sample data set 50 the results are best i.e. 85%.

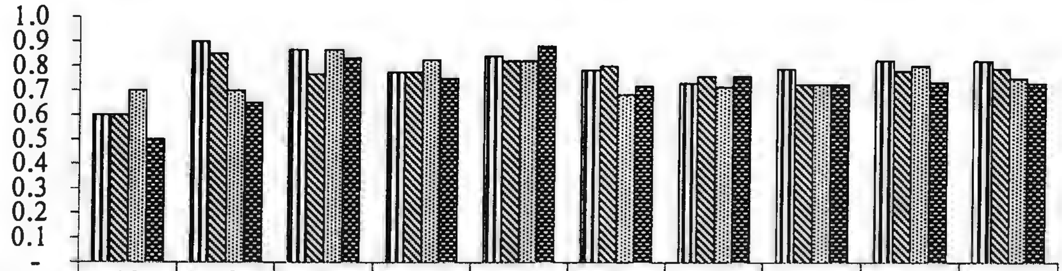


	10	20	30	40	50	60	70	80	90	100
▥ SVM	0.60	0.90	0.87	0.85	0.88	0.80	0.75	0.80	0.85	0.83
▧ NN	0.60	0.85	0.77	0.78	0.83	0.81	0.76	0.73	0.78	0.80
▤ NB	0.71	0.70	0.87	0.85	0.83	0.74	0.73	0.76	0.82	0.76
▦ BN	0.50	0.65	0.84	0.75	0.88	0.72	0.76	0.73	0.74	0.73

4-26: Model Wise Precision Analysis Using AR (All Features)

Result: Model Wise Recall Analysis of Features – All Features

Similarly figure 4-27 show the recall result of features using classification models provide mostly the same result it gives in its precision in figure 7. Figure 8 shows that SVM Model gives better result i.e. 79%.

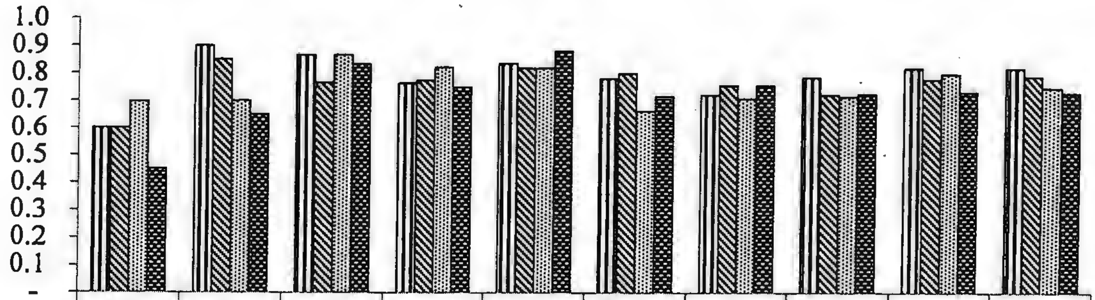


	10	20	30	40	50	60	70	80	90	100
□ SVM	0.60	0.90	0.87	0.78	0.84	0.78	0.73	0.79	0.82	0.82
▨ NN	0.60	0.85	0.77	0.78	0.82	0.80	0.76	0.73	0.78	0.79
▦ NB	0.70	0.70	0.87	0.83	0.82	0.68	0.71	0.73	0.80	0.75
▩ BN	0.50	0.65	0.83	0.75	0.88	0.72	0.76	0.73	0.73	0.73

4-27: Model Wise Recall Analysis Using AR (All Features)

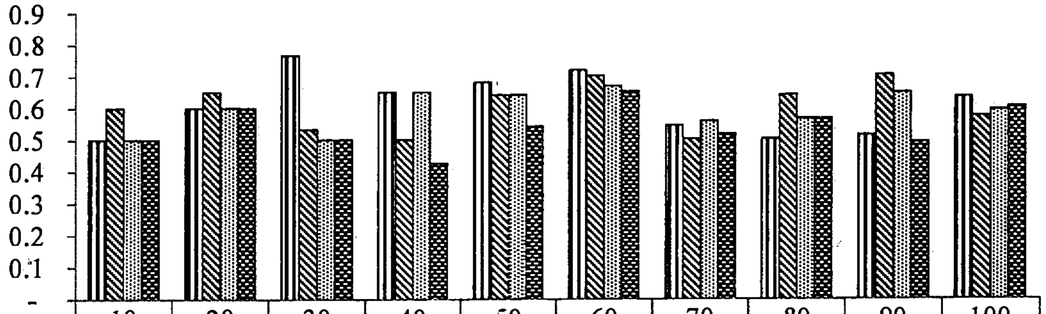
Result: Model Wise F-Measure Analysis of Features – All Features

Similarly, in figure 4-28 the f-measure result of features using classification models shows that SVM classifier outperforms with accuracy of 79%. The best average results are on sample data set of 50 i.e. 84% similar to precision and recall calculations. Likewise, the worst outcome is with smallest data set of 10 i.e. 59%.



	10	20	30	40	50	60	70	80	90	100
□ SVM	0.60	0.90	0.87	0.76	0.84	0.78	0.72	0.79	0.82	0.82
▨ NN	0.60	0.85	0.77	0.77	0.82	0.80	0.76	0.72	0.78	0.79
▦ NB	0.70	0.70	0.87	0.82	0.82	0.66	0.71	0.72	0.80	0.75
▩ BN	0.45	0.65	0.83	0.75	0.88	0.72	0.76	0.73	0.73	0.73

4-28: Model Wise F-Measure Analysis Using AR (All Features)

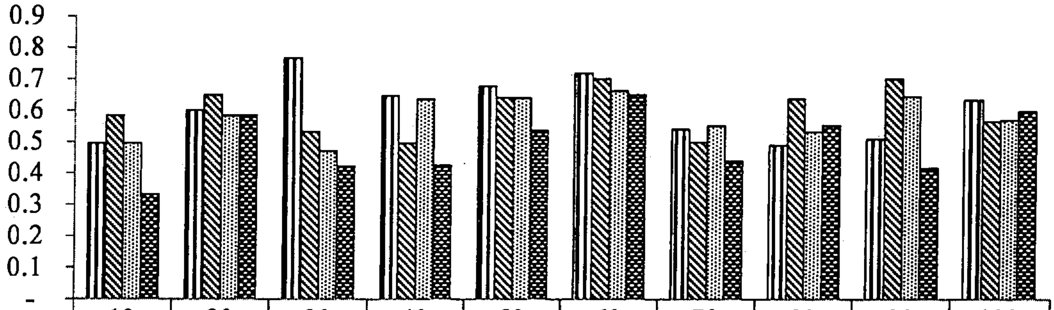


	10	20	30	40	50	60	70	80	90	100
□SVM	0.50	0.60	0.77	0.65	0.68	0.72	0.54	0.50	0.51	0.63
▨NN	0.60	0.65	0.53	0.50	0.64	0.70	0.50	0.64	0.70	0.57
▦NB	0.50	0.60	0.50	0.65	0.64	0.67	0.56	0.56	0.64	0.59
▩BN	0.50	0.60	0.50	0.43	0.54	0.65	0.51	0.56	0.49	0.60

4-30: Model Wise Recall Analysis Using ARIR (All Features)

Result: Model Wise F-Measure Analysis of Features – All Features

The figure 4-31 shows f-measure analysis on 2nd data set i.e. ARIR. The SVM model has produced best results of weighted average of 61%. The Neural Net classifier has shown 2nd best results of 60% accuracy. The weighted average of all models produces best of 68% results on sample data set of size 60. The sample data set 10 carries worst average results of 48%.



	10	20	30	40	50	60	70	80	90	100
□SVM	0.50	0.60	0.77	0.65	0.68	0.72	0.54	0.49	0.51	0.63
▨NN	0.58	0.65	0.53	0.50	0.64	0.70	0.50	0.64	0.70	0.56
▦NB	0.50	0.58	0.47	0.64	0.64	0.66	0.55	0.53	0.64	0.57
▩BN	0.33	0.58	0.42	0.43	0.54	0.65	0.44	0.55	0.42	0.59

4-31: Model Wise F-Measure Analysis Using ARIR (All Features)

4.3.7 Model Wise Combined Features Analysis [Selected Features]

Each classifier is executed against selected features on randomly selected sample data sets of 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 for AR and ARIR.

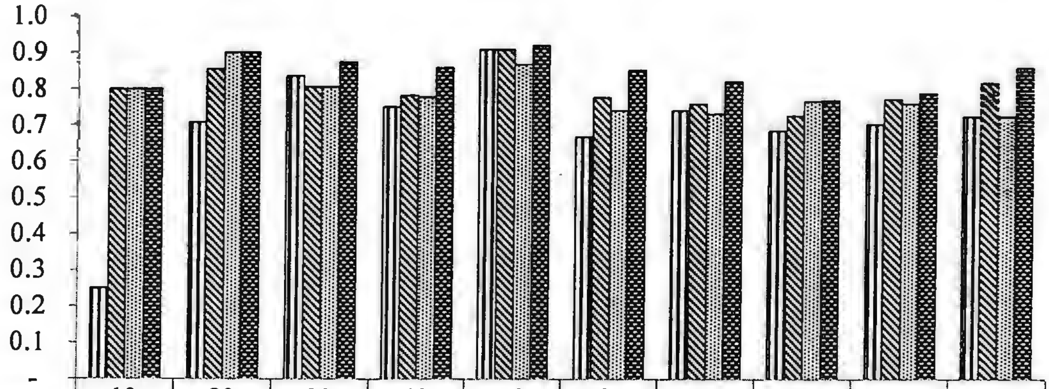
Scenario 1: Combined Features Analysis Based on Average Revenue (AR) – Selected Features

In this section we analyzed precision, recall and f-measures of selected features COR, RCP, RCR, CDN, CFR and CDR by using classification models on first type of datasets i.e. AR.

Result: Model Wise Precision Analysis of Features – Selected Features

Figure 4-32 show the precision results of selected features by using four different classification models. The comparison of results between different classification models shows the performance of SVM model is better than other models. The SVM model offers the highest average accuracy result of 84%. The SVM model outperforms with 92% on data set of 50 instances. The NB classifier show 2nd best average precision of 80%.

The sample data set 10 carries worst average results of 66% and on sample data set 50 the results are best i.e. 90%.

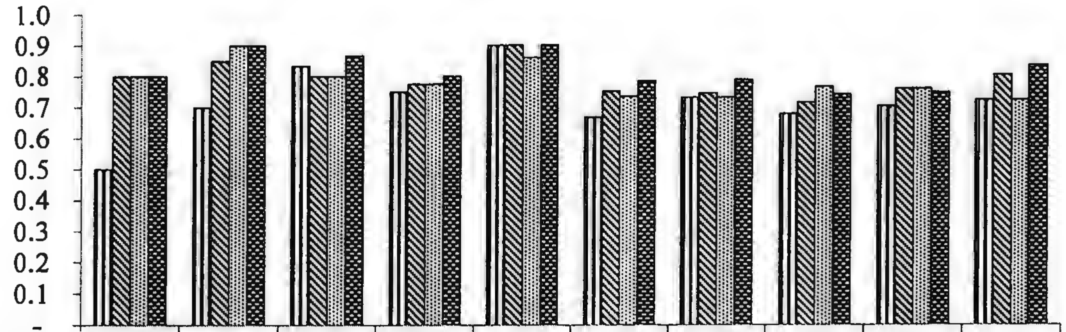


	10	20	30	40	50	60	70	80	90	100
□ BN	0.25	0.71	0.84	0.75	0.91	0.67	0.74	0.68	0.70	0.72
▧ NB	0.80	0.85	0.81	0.78	0.91	0.78	0.76	0.72	0.77	0.81
▦ NN	0.80	0.90	0.81	0.78	0.87	0.74	0.73	0.76	0.76	0.72
▩ SVM	0.80	0.90	0.87	0.86	0.92	0.85	0.82	0.77	0.79	0.85

4-32: Model Wise Precision Analysis Using AR (Selected Features)

Result: Model Wise Recall Analysis of Features – Selected Features

Similarly figure 4-33 show the recall result of selected features using classification models provide mostly the same result it gives in its precision in figure 7. Figure 8 shows that SVM Model gives better result i.e. 81% and the best average accuracy of selected features against four classifiers is 89% on data set size 50.

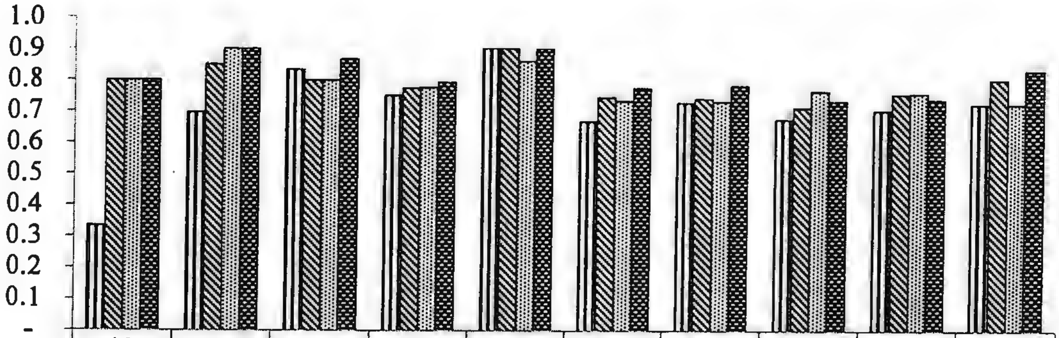


	10	20	30	40	50	60	70	80	90	100
□ BN	0.50	0.70	0.83	0.75	0.90	0.67	0.73	0.68	0.70	0.72
▧ NB	0.80	0.85	0.80	0.78	0.90	0.75	0.74	0.71	0.76	0.80
▦ NN	0.80	0.90	0.80	0.78	0.86	0.73	0.73	0.76	0.76	0.72
▩ SVM	0.80	0.90	0.87	0.80	0.90	0.78	0.79	0.74	0.74	0.83

4-33: Model Wise Recall Analysis Using AR (Selected Features)

Result: Model Wise F-Measure Analysis of Features– Selected Features

Similarly, in figure 4-34 the f-measure result of selected features (COR, RCP, RCR, CDN, CFR, CDR) using classification models shows that SVM classifier outperforms with accuracy of 81%. The best average results are on sample data set of 50 i.e. 89% similar to precision and recall calculations. Likewise, the worst outcome is with smallest data set of 10 i.e. 68%.



	10	20	30	40	50	60	70	80	90	100
□ BN	0.33	0.70	0.83	0.75	0.90	0.67	0.73	0.67	0.70	0.72
▧ NB	0.80	0.85	0.80	0.77	0.90	0.74	0.74	0.71	0.75	0.80
▦ NN	0.80	0.90	0.80	0.78	0.86	0.73	0.73	0.76	0.76	0.72
▩ SVM	0.80	0.90	0.87	0.79	0.90	0.77	0.78	0.73	0.74	0.83

4-34: Model Wise F-Measure Analysis Using AR (Selected Features)

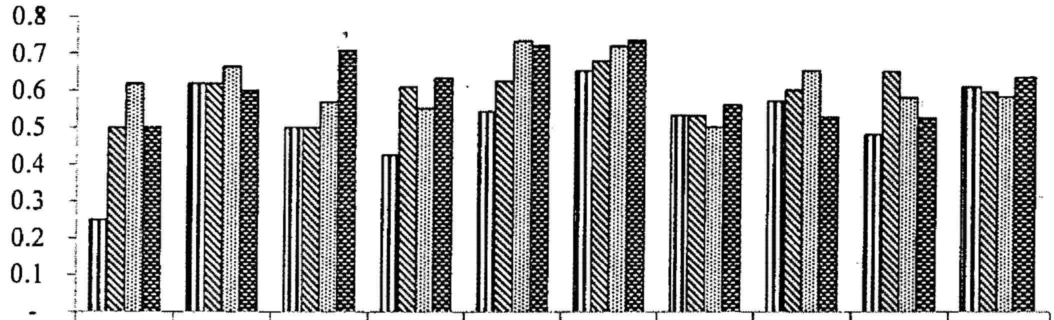
Scenario 2: Combined Features Analysis Based on Average Relative Increase in Revenue (ARIR) – Selected Features

In this section we analyze model wise precision, recall and f-measure on range of ARIR data sets (10 to 100) having selected features (SCW excluded).

Result: Model Wise Precision Analysis of Features – Selected Features

Figure 4-35 show the precision results of features by using four different classification models. The comparison of results between different classification models shows that Neural Net classifier outperforms equally among all classifiers with weighted average of 62%. SVM Model gives the highest average accuracy result of 73% on data set of 60 instances. After Neural Net, the SVM classifier has produced results of accuracy 61%.

The sample data set 10 carries worst average results of 47% and on sample data set 60 the results are best i.e. 70%. This is also evident that AR results are better than ARIR.

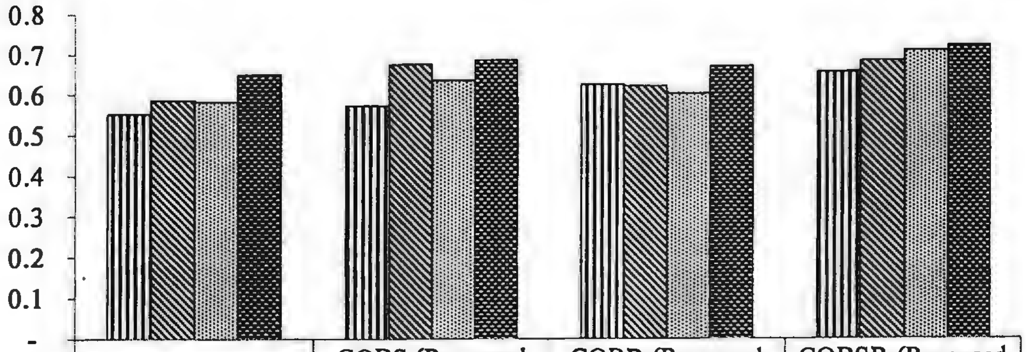


	10	20	30	40	50	60	70	80	90	100
□ BN	0.25	0.62	0.50	0.43	0.54	0.65	0.53	0.57	0.48	0.61
▧ NB	0.50	0.62	0.50	0.61	0.63	0.68	0.53	0.60	0.65	0.59
▤ NN	0.62	0.67	0.57	0.55	0.73	0.72	0.50	0.65	0.58	0.58
▩ SVM	0.50	0.60	0.71	0.63	0.72	0.73	0.56	0.53	0.52	0.63

4-35: Model Wise Precision Analysis Using ARIR (Selected Features)

Result: Model Wise Recall Analysis of Features – Selected Features

Figure 4-36 show the recall results of features of 2nd dataset (ARIR) Average Relative Increase in revenue by using four classification models. In figure 11, we observed that models SVM and NN gives best average result of accuracy 61%. Next to SVM and NN mode, the NB classifier has produced the performance of 58%. The SVM classifier shows best results of 73% on sample set 60. The weighted average of all models produces best of 69% results on sample data set of size 60.

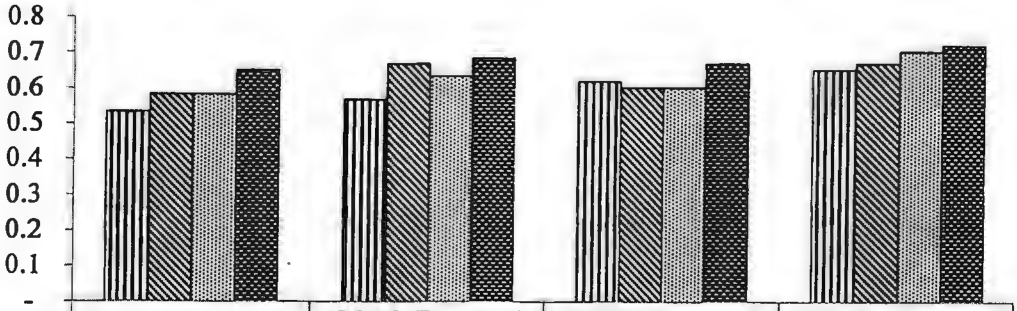


	Co-BM (Baseline)	COBS (Proposed Method-I)	COBR (Proposed Method-II)	COBSR (Proposed Method-III)
▯BN	0.55	0.57	0.62	0.65
▧NB	0.59	0.67	0.62	0.68
▦NN	0.58	0.63	0.60	0.70
▩SVM	0.65	0.68	0.67	0.72

4-41: Precision of Baseline Analysis Using ARIR (All Features)

Result: Recall of Baseline Analysis (ARIR) – All Features

The figure 4-42 shows the precision results against baseline analysis by using classification models on best sample data set of 60 instances. The SVM classifier outperforms with average accuracy of 68%. Whereas, the COBSR feature set has produced the best average results of 68% that are better than baseline Co-BM feature set having 59%.

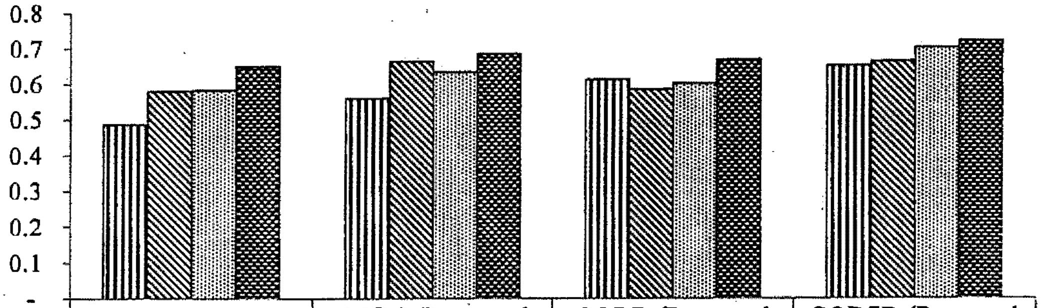


	Co-BM (Baseline)	COBS (Proposed Method-I)	COBR (Proposed Method-II)	COBSR (Proposed Method-III)
▯BN	0.53	0.57	0.62	0.65
▧NB	0.58	0.67	0.60	0.67
▦NN	0.58	0.63	0.60	0.70
▩SVM	0.65	0.68	0.67	0.72

4-42: Recall of Baseline Analysis Using ARIR (All Features)

Result: F-Measure of Baseline Analysis (ARIR) – All Features

The figure 4-43 shows the precision results against baseline analysis by using classification models on best sample data set of 60 instances. The SVM classifier outperforms with average accuracy of 68%. Whereas, the COBSR feature set has produced the best average results of 68% that are better than baseline Co-BM feature set having 58%.



	Co-BM (Baseline)	COBS (Proposed Method-I)	COBR (Proposed Method-II)	COBSR (Proposed Method-III)
□BN	0.49	0.56	0.61	0.65
▧NB	0.58	0.66	0.58	0.66
▤NN	0.58	0.63	0.60	0.70
▩SVM	0.65	0.68	0.67	0.72

4-43: F-Measure of Baseline Analysis Using ARIR (All Features)

4.3.9 Baseline Analysis - Selected Features

For baseline analysis against selected features, we considered the results of Co-BM selected feature set as baseline and three solutions are proposed (COBS, COBR, COBSR) and compared with baseline. All four models Neural Net, SVM, Bayes Net and Naïve Bayes are applied on AR and ARIR best data sets.

Table 12: Baseline Analysis (AR) – Selected Features

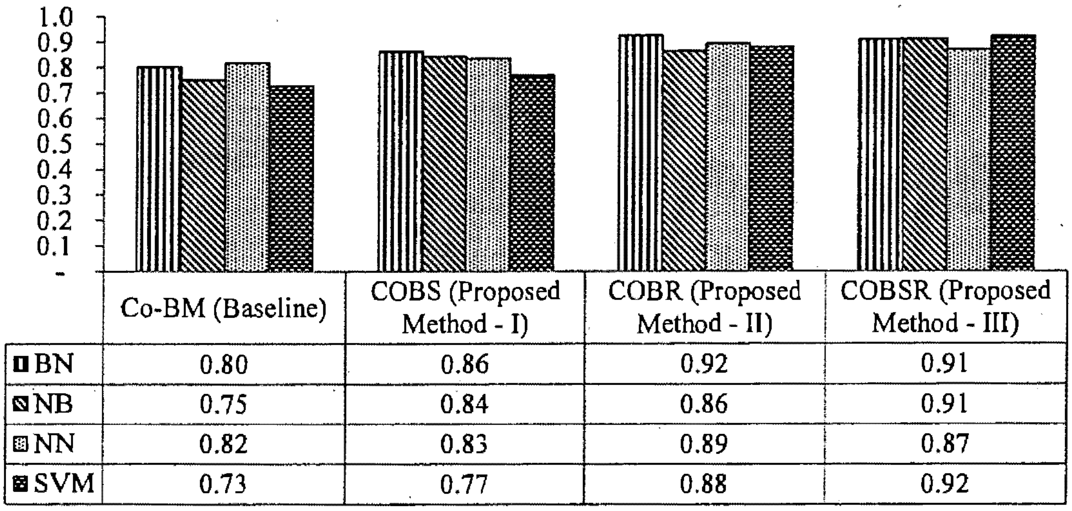
COBS	On-Time Rectification	Faults Registered	Disconnection
	Customer Denied Rectification		
COBR	On-Time Rectification	Faults Registered	Disconnection
	Customer Denied Rectification	Collection Ratio	Capacity
COBSR	On-Time Rectification	Faults Registered	Disconnection
	Customer Denied Rectification	Collection Ratio	Capacity

Scenario 1: Baseline Analysis Based on Average Revenue (AR) - Selected Features

In this section, we evaluated baseline analysis against selected features to determine precision, recall and f-measure of AR sample sets.

Result: Precision of Baseline Analysis (AR) - Selected Features

The figure 4-44 shows the precision results against baseline analysis by using classification models on best sample data set of 50 instances against selected feature subset. The proposed method III (COBSR) has produced the best average results of 90%. The SVM classifier outperforms in proposed method III (COBSR) with accuracy of 92%. Whereas, with each proposed method the result gets improved from 77% to 90%. Against each proposed method, the generative models (NB, BN) produces best results because of high info gain and gain ratio.

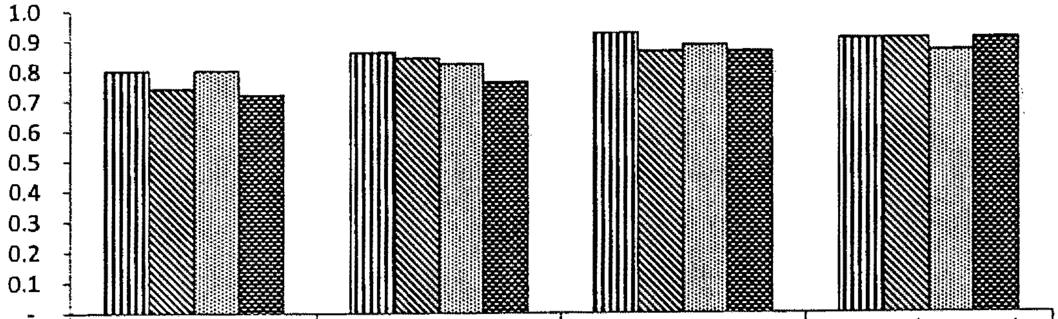


	Co-BM (Baseline)	COBS (Proposed Method - I)	COBR (Proposed Method - II)	COBSR (Proposed Method - III)
□ BN	0.80	0.86	0.92	0.91
▧ NB	0.75	0.84	0.86	0.91
▦ NN	0.82	0.83	0.89	0.87
▩ SVM	0.73	0.77	0.88	0.92

4-44: Precision of Baseline Analysis Using AR (Selected Features)

Result: Recall of Baseline Analysis (AR) - Selected Features

The figure 4-45 shows the recall results against baseline analysis by using classification models on best sample data set of 50 instances against selected feature subset. The proposed method III (COBSR) has produced the best average results of 90%. The SVM, BN and NB classifier outperforms in proposed method III (COBSR) with same accuracy of 90%. Whereas, with each proposed method the result gets improved from 77% to 89%.

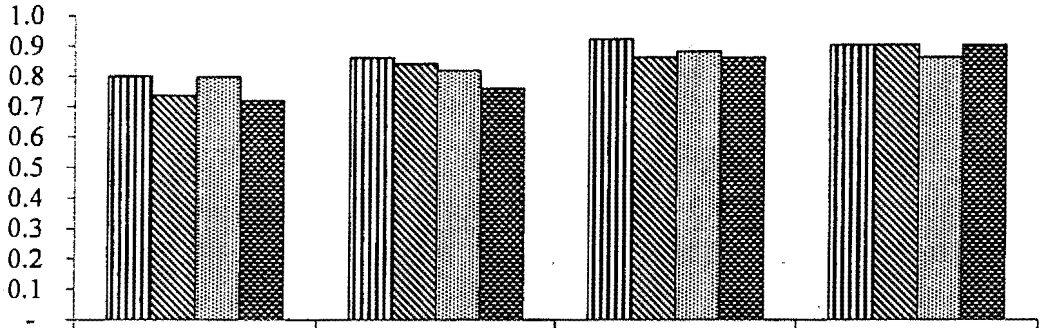


	Co-BM (Baseline)	COBS (Proposed Method - I)	COBR (Proposed Method - II)	COBSR (Proposed Method - III)
□ BN	0.80	0.86	0.92	0.90
▨ NB	0.74	0.84	0.86	0.90
▦ NN	0.80	0.82	0.88	0.86
▩ SVM	0.72	0.76	0.86	0.90

4-45: Recall of Baseline Analysis Using AR (Selected Features)

Result: F-Measure of Baseline Analysis (AR) - Selected Features

The figure 4-46 shows the f-measure results against baseline analysis by using classification models on best sample data set of 50 instances against selected feature subset. The proposed method III (COBSR) has produced the best average results of 89%. The SVM, BN and NB classifier outperforms in proposed method III (COBSR) with same accuracy of 90%.



	Co-BM (Baseline)	COBS (Proposed Method - I)	COBR (Proposed Method - II)	COBSR (Proposed Method - III)
□ BN	0.80	0.86	0.92	0.90
▨ NB	0.74	0.84	0.86	0.90
▦ NN	0.80	0.82	0.88	0.86
▩ SVM	0.72	0.76	0.86	0.90

4-46: F-Measure of Baseline Analysis Using AR (Selected Features)

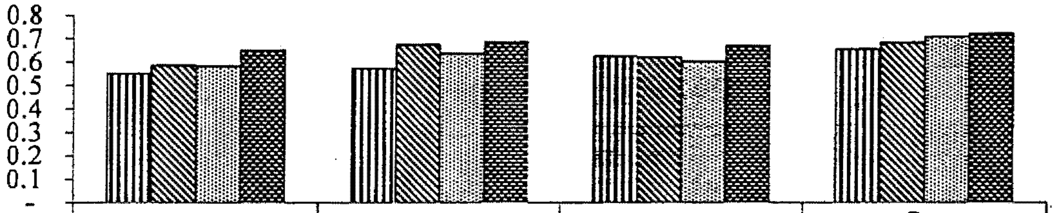
Whereas, with each proposed method the result gets improved from 76% to 89%. Since BN and NB classifiers mainly relay on info gain and gain ratio of individual features so with selected features in each proposed method these models outperform.

Scenario 2: Baseline Analysis Based on Average Relative Increase in Revenue (ARIR) - Selected Features

In this section, we evaluated baseline analysis against selected features of ARIR sample data set of size 50 to determine precision, recall and f-measure.

Result: Precision of Baseline Analysis (ARIR) - Selected Features

The figure 4-47 shows the precision results against baseline analysis by using classification models on best sample data set of 60 instances against selected features. The SVM classifier outperforms with average accuracy of 68%. Whereas, the COBSR feature set has produced the best average results of 70% that are better than baseline Co-BM feature set having 59%. Similarly, other two proposed methods have also produced better average results as compared to baseline Co-BM features.

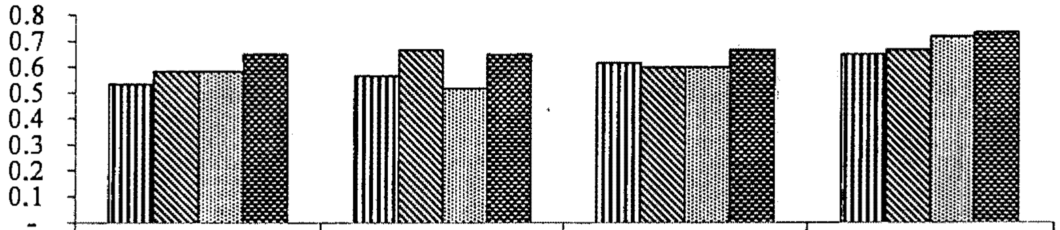


	Co-BM (Baseline)	Co-BM+SBM (Proposed Method-I)	Co-BM+RGM (Proposed Method-II)	Co-BM+SBM+RGM (Proposed Method-III)
▯ BN	0.55	0.57	0.62	0.65
▧ NB	0.59	0.67	0.62	0.68
▤ NN	0.58	0.63	0.60	0.70
▩ SVM	0.65	0.68	0.67	0.72

4-47: Precision of Baseline Analysis Using ARIR (Selected Features)

Result: Recall of Baseline Analysis (ARIR) - Selected Features

The figure 4-48 shows the recall results against baseline analysis by using classification models on best sample data set of 60 instances against selected features. The SVM classifier outperforms with average accuracy of 68%. Whereas, the COBSR feature set has produced the best average results of 69% that are better than baseline Co-BM feature set having 59%. Similarly, other two proposed methods have also produced better average results as compared to baseline Co-BM features.

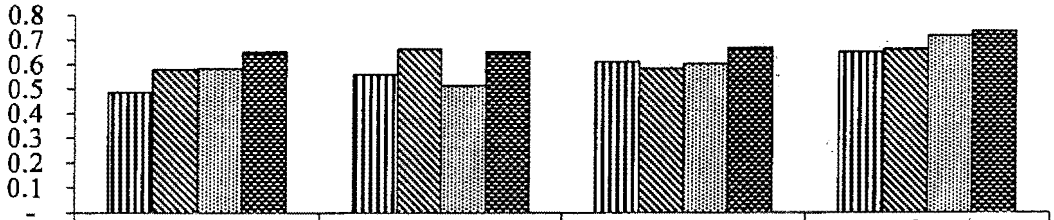


	Co-BM (Baseline)	Co-BM+SBM (Proposed Method-I)	Co-BM+RGM (Proposed Method-II)	Co-BM+SBM+RGM (Proposed Method-III)
▯BN	0.53	0.57	0.62	0.65
▧NB	0.58	0.67	0.60	0.67
▤NN	0.58	0.52	0.60	0.72
▩SVM	0.65	0.65	0.67	0.73

4-48: Recall of Baseline Analysis Using ARIR (Selected Features)

Result: F-Measure of Baseline Analysis (ARIR) - Selected Features

The figure 4-49 shows the f-measure results against baseline analysis by using classification models on best sample data set of 60 instances against selected features. The SVM classifier outperforms with average accuracy of 67%. Whereas, the COBSR feature set has produced the best average results of 69% that are better than baseline Co-BM feature set having 58%. Similarly, other two proposed methods have also produced better average results as compared to baseline Co-BM features.

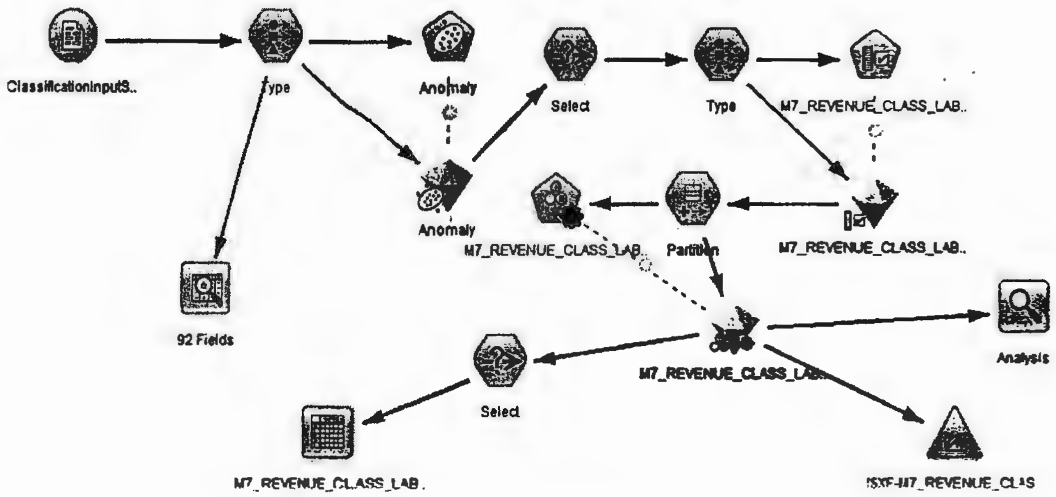


	Co-BM (Baseline)	Co-BM+SBM (Proposed Method-I)	Co-BM+RGM (Proposed Method-II)	Co-BM+SBM+RGM (Proposed Method-III)
▯BN	0.49	0.56	0.61	0.65
▧NB	0.58	0.66	0.58	0.66
▤NN	0.58	0.51	0.60	0.72
▩SVM	0.65	0.65	0.67	0.73

4-49: F-Measure of Baseline Analysis Using ARIR (Selected Features)

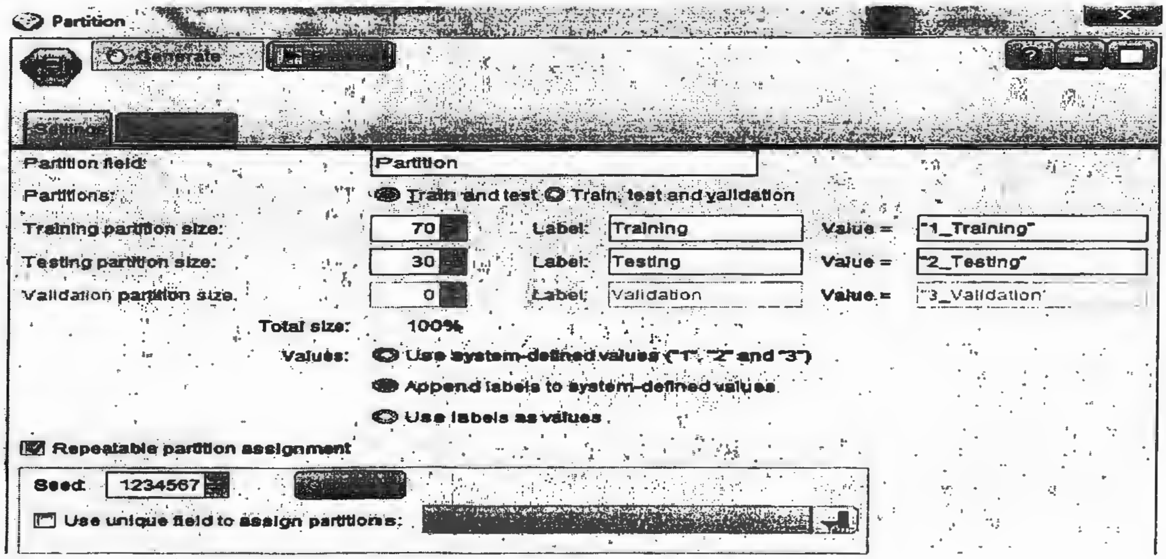
4.3.10 Testing the Model

All instances of six months' data from Sep 2014 to Feb 2015 are evaluated for prediction wherein the class label is an average of all month's revenue i.e from Sep 2014 to Sep 2015. A model has been built in IBM SPSS Modeler as in snapshot 4-50. The anomaly / outliers is identified and then the expert modeler is applied on the data set.



4-50: IBM SPSS Model

The model is trained on partition size of 70% and tested on 30% of the partition as in figure 4.51.



Partition field: Partition

Partitions: ● Train and test ○ Train, test and validation

Training partition size: 70 Label: Training Value = "1_Training"

Testing partition size: 30 Label: Testing Value = "2_Testing"

Validation partition size. 0 Label: Validation Value = "3_Validation"

Total size: 100%

Values: ○ Use system-defined values ("1", "2" and "3")

● Append labels to system-defined values

○ Use labels as values

☑ Repeatable partition assignment

Seed 1234567

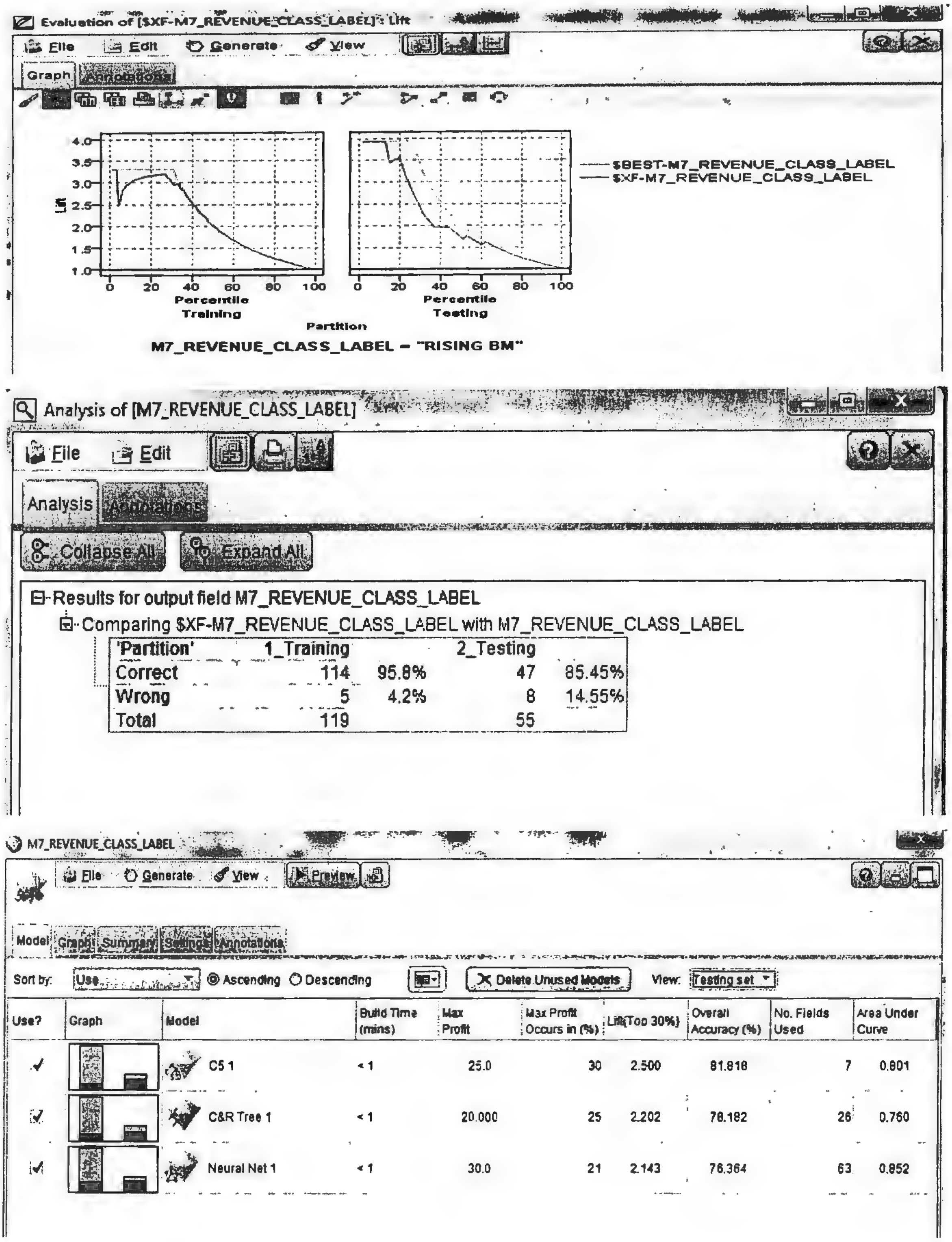
□ Use unique field to assign partitions:

4-51: IBM SPSS Training & Testing Snapshot

All following models filtered in expert modeler are applied.

1. C5
2. Logistic regression
3. Decision list
4. Bayesian Network
5. Discriminant
6. LSVM
7. Random Trees
8. Tree-AS
9. Neural Net
10. C&R Tree
11. Quest
12. CHAID

The maximum lift is produced by model C5, C&R Tree and Neural Net as 2.5, 2.2 and 2.1 respectively and overall accuracy of testing set is between 82 to 76%.



Results for output field M7_REVENUE_CLASS_LABEL

Comparing $XF-M7_REVENUE_CLASS_LABEL with M7_REVENUE_CLASS_LABEL

'Partition'	1_Training		2_Testing	
Correct	114	95.8%	47	85.45%
Wrong	5	4.2%	8	14.55%
Total	119		55	

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
✓		C5 1	< 1	25.0	30	2.500	81.818	7	0.801
✓		C&R Tree 1	< 1	20.000	25	2.202	78.182	26	0.760
✓		Neural Net 1	< 1	30.0	21	2.143	76.364	63	0.852

4-52: IBM SPSS Model Results

The training results are 96% correct and testing results are 85% correct.

4.3.11 ROC Curve

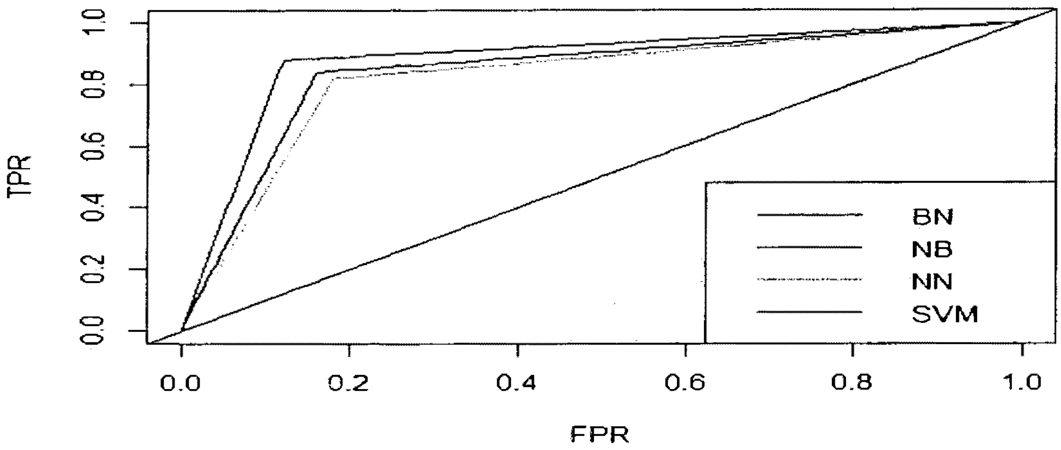
The receiver operating characteristic (ROC) is a curve of true positive rate (TPR) against false positive rate (FPR). The TPR is known as sensitive or recall whereas FPR is called fall-out and is calculated as 1-specificity. The ROC curve is plotted by applying classifiers on best data set size having all features.

ROC Curve on AR Data Set

Table 13: ROC Values (AR)

Classifier	Data Size	Method	TP Rate	FP Rate
BN	50	AR	0.880	0.120
NB	50	AR	0.820	0.180
NN	50	AR	0.820	0.180
SVM	50	AR	0.840	0.160

ROC plot on AR Data Set



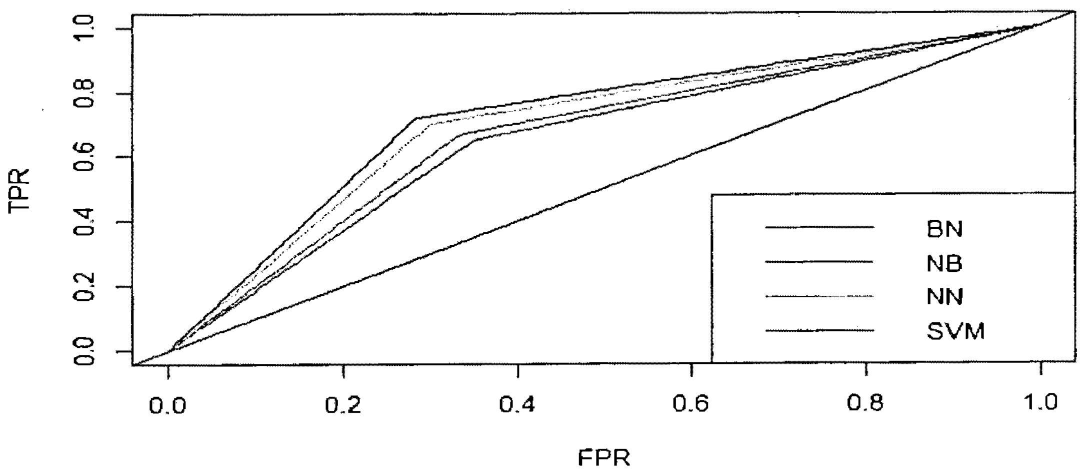
4-53: ROC Curves (AR)

ROC Curve on ARIR Data Set

Table 14: ROC Values (ARIR)

Classifier	Data Size	Method	TP Rate	FP Rate
BN	60	ARIR	0.650	0.350
NB	60	ARIR	0.667	0.333
NN	60	ARIR	0.700	0.300
SVM	60	ARIR	0.717	0.283

ROC plot on ARIR Data Set



4-54: ROC Curves (ARIR)

We have predicted the rising star based on ranking derived from manager's feature set and then compared it with yearly rating of business managers. This technique is applied on both scenarios i.e. AR and ARIR.

4.3.12 Rising Star Score (AR)

The rising star is determined based on score derived on feature set as per following formulae.

Rising Star Score (AR) = COP + COR + CAN + SBR + SCR + SCW ÷ RCS + RCR + RCP – (CFR + CRP + CRT + CDN + CDR + CDI)

Since features like faults, disconnection and customer denial carries indirect relation that is lower is better so we subtracted the sum of negative features score from positive feature score.

Table 15: Features Direct Inverse Relation

COP	Co-BM	On-Time Provisioning	Direct Relation, Positive
COR	Co-BM	On-Time Rectification	Direct Relation, Positive
CAN	Co-BM	Net Adds	Direct Relation, Positive
SBR	SBM	BB To PSTN Ratio	Direct Relation, Positive
SCR	SBM	Customer Retention	Direct Relation, Positive
SCW	SBM	Customer Winback	Direct Relation, Positive
RCS	RGM	Customer Segmentation	Direct Relation, Positive

RCR	RGM	Collection Ratio	Direct Relation, Positive
RCP	RGM	Capacity	Direct Relation, Positive
CFR	Co-BM	Faults Registered	Indirect Relation, Negative
CRP	Co-BM	Repeat Faults	Indirect Relation, Negative
CRT	Co-BM	Repeat Telephone	Indirect Relation, Negative
CDN	Co-BM	Disconnection	Indirect Relation, Negative
CDR	Co-BM	Customer Denied Rectification	Indirect Relation, Negative
CDI	Co-BM	Customer Denied Installation	Indirect Relation, Negative

Top 10 Business Manager - AR

In table 16, the net progress is rising star score and business score is score calculated by business based on weightage and features relationship of AR data set. The Top 10 BMs are more or less in top 15 as per business score also.

Table 16: Top 10 BMs (Rising Business Managers) – AR

	BM ID	**Net Progress**	**Business Score**
1	10050002	515.26	7
2	10062849	506.6	3
3	10025555	415.48	2
4	10062184	400.49	8
5	10052248	399.61	14
6	10049420	398.34	10
7	10062021	394.36	13
8	10034582	393.24	12
9	10028181	392.71	15
10	10014847	389.33	16

Bottom 10 Business Manager – AR

In Table 17, the net progress is rising star score and business score is score calculated by business based on weightage and features relationship of AR data set. The bottom 10 business managers are in worst performance segment as per business score also.

Table 17: Bottom 10 BMs (Rising Business Managers) – AR

	BM ID	**Net Progress**	**Business Score**
166	10035228	218.41	170
167	10034673	217.51	169
168	10062846	217.06	133
169	10062004	203.25	108
170	10049808	193.58	167
171	10028419	192.59	168

172	10017071	134.38	173
173	10027251	108.47	172
174	10039735	4.14	174
175	10062855	0	175

Whereas, table 18 represents the correlation of business managers w.r.t. order of the business managers.

Table 18: Output Correlation

Net Progress Order	Business Score Order	Revenue Sequence (Avg.)	Revenue Sequence (last standing)
10050002	10028585	10049460	10062891
10062849	10025555	10062891	10049460
10025555	10062849	10017793	10017793
10062184	10027962	10028622	10021037
10052248	10060485	10007265	10007265
10049420	10025423	10050365	10028622
10062021	10050002	10021037	10050365
10034582	10062184	10025707	10062673
10028181	10035614	10021895	10025707
10014847	10049420	10062673	10021895
PearsonsCorrelation	-0.46	-0.0307	0.255216
KendallsCorrelation	-0.377778	0.1111	0.1555555

4.3.13 Rising Star (ARIR)

The rising star is determined based on score derived on feature set as per following formulae. Rising Star Score (ARIR) = COP + COR + CAN + SBR + SCR + SCW + RCS + RCR + RCP – (CFR + CRP + CRT + CDN + CDR + CDI)

Since features like faults, disconnection and customer denial carries indirect relation that is lower is better so we subtracted the sum of negative features score from positive feature score.

Top 10 Business Manager – ARIR

In table 19, the net progress is rising star score and business score is score calculated by business based on weightage and features relationship of ARIR data set. The Top 10 BMs are more or less in top 15 as per business score also.

Table 19: Top 10 Rising Business Managers (ARIR)

	BM ID	Net Progress	Business Score Rate
1	10062849	490.1	3
2	10050002	425.27	7
3	10028181	400.9	15
4	10049420	394.15	10
5	10050399	383.09	55
6	10034582	378.59	12
7	10062021	375.85	13
8	10021016	371.87	105
9	10062383	371.34	89
10	10031127	371.17	35

Bottom 10 Business Manager – ARIR

In Table 20, the net progress is rising star score and business score is score calculated by business based on weightage and features relationship of ARIR data set. The bottom 10 business managers are in worst performance segment as per business score also.

Table 20: Bottom 10 Not Rising Business Managers (ARIR)

	BM ID	Net Progress	Business Score Rate	Revenue base Sequence
166	10035228	242.35	170	159
167	10063798	238.9	145	110
168	10021893	220.62	121	49
169	10021891	219.41	155	111
170	10028201	218.19	129	86
171	10028419	192.55	168	140
172	10017071	167.44	173	158
173	10027251	151.1	172	173
174	10039735	8.82	174	174
175	10062855	0.1	175	175

CHAPTER 5

CONCLUSION AND FUTURE WORK

5. Conclusion and Future Work

The prediction of rising business manager based on performance of co-business managers is a very crucial business requirement for deputation and replacement. Based on machine learning algorithms, we compared the prediction results with business current standing scores. We proposed three methods and compared the results with baseline co-BM results. The discriminative classifiers particularly SVM produces best results beside neural net because the features correlation is not linearly separable, hence require vector support.

In future work, we will scale this effort to multiple class labels and will use the deep learning classifiers such as deep belief network to find better results.

6. References

[1] C. H. J. T. Y. Z. X. Rui Yan, "To Better Stand on the Shoulder of Giants," in JCDL, Washington, DC, USA, 2012.

[2] R. A. a. F. M. Ali Daud, ""Finding Rising Stars in Social Networks"," in The International Conference on Database Systems for Advanced Applications (DASFAA), 2013.

[3] C. S. F. K. L. T. S.-K. N. Xiao-Li Li, ""Searching for Rising Stars in Bibliography Networks"," in The International Conference on Database Systems for Advanced Applications (DASFAA), 2009.

[4] E. A. N. Qasem A. Al-Radaideh, ""Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance"," in International Journal of Advanced Computer Science and Applications (IJACSA), 2012.

[5] L. V. Fabrico Voznika, "http://courses.cs.washington.edu/courses," 10-05-2015.

[6] M. K. J. P. Jiawei Han, ""Data Mining Concepts and Techniques"," in Morgan Kaufmann Publishers, 2011.

[7] M. A. M. M. D. C. Ali Daud, ""Using Machine Learning Techniques for Rising Star Prediction in Co-author Network"," in Scientometrics, 2014.

[8] SVMWikipedia, "http://en.wikipedia.org/wiki/Support_vector_machine," 10-05-2015.

[9] LandlineWikipedia, "http://en.wikipedia.org/wiki/Landline," 10-05-2015.

[10] F. X. W. W. X. B. S. Y. T. M. B. Jun Zhang, "CocaRank: A Collaboration Caliber-based Method for Finding Academic Rising Stars," in International World Wide Web Conference Committee, Montréal, Québec, Canada, 2016.

[11] G. Michael J., ""Examining IT Professionals Adaptation to Technological Change:The Influence of Gender andPersonal Attributes"," in The DATA BASE for Advances in Information Systems, 2004.

[12] A. R. H. Z. A. O. Hamidah Jantan, ""Classification and prediction of academic talent using data mining techniques"," in International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2009.

[13] A. R. H. Z. A. O. Hamidah Jantan, ""Human Talent Forecasting using Data Mining Classification Techniques"," in International Journal of Technology Diffusion, 2010.

[14] W. Z. R. B. Xiaobo Xu, ""IT infrastructure capabilities and IT project success: a development team perspective"," in Information Technology and Management, 2010.

[15] L. C. Chenfu Chien, ""Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry"," in Expert Systems with Applications: An International Journal, 2008.

[16] C. Ebert, ""The impacts of software product management"," in The Journal of Systems and Software, 2007.

[17] G. V. I. &. N. Tsatsaronis, ""How to Become a Group Leader? Or Modeling Author Types Based on Graph"," in The International Conference on Theory and Practice of Digital Libraries (TPDL), 2011.

[18] L. Z. C. M. R. M. A. H. R. S. Dewan Md. Farid, ""Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks"," in Expert Systems with Applications: An International Journal, 2014.

[19] C.-C. C. F.-Y. C. Ling-Jing Kao, ""A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring"," in Knowledge-Based Systems, 2012.

[20] J. S. J. W. Hui Li, "Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods," in Expert Systems with Applications: An International Journal, 2010 .

[21] O. M. Lior Rokach, ""Data Mining With Decision Trees: Theory and Applications"," in World Scientific Publishing Co., 2014.

[22] T. M. Mitchell, "GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION," in Machine Learning, 2010, pp. 1-17.

[23] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2007.

[24] I. H. Witten and M. A. H. Eibe Frank, Data Mining: Practical machine learning tools and techniques, 3rd Edition, San Francisco: Morgan Kaufmann, 2011.