Default of MasterCard Customer Prediction Using Machine Learning Approaches

Thesis Report

submitted in partial fulfillment of the requirements
for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Vaishali
(801532057)

Under the supervision of:

Dr. Rajkumar Tekchandani
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004
June 2017

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "**Default Of MasterCard Customer Prediction Using Machine Learning Approaches**", in partial fulfilment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of **Mr. Rajkumar Tekchandani** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Vaishali
801532057
ME - CSE

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Mr. Rajkumar Tekchandani
Assistant Professor
Computer Science and Engineering Department
Thapar University, Patiala

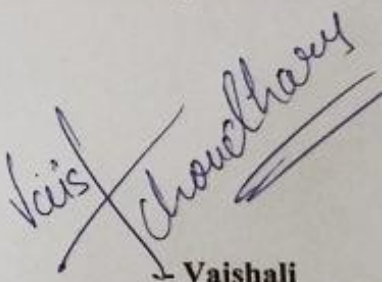
Acknowledgement

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds. With the profound sense of gratitude and heartiest regard, I express my sincere feelings of indebtedness to my guide **Mr. Rajkumar Tekchandani**, Assistant Professor, Computer Science and Engineering Department, Thapar University for his positive attitude, constant encouragement, keen interest, invaluable cooperation, generous attitude and above all her blessings. He has been a source of inspiration for me.

I am grateful to **Dr. Maninder Singh**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar University for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academics Affairs in the University for making provisions of infrastructure such as library facilities, computer labs equipped with internet facility, immensely useful for the learners to equip themselves with latest in the field.

Last but not the least I would like to express my heartfelt thanks to my parents and my friends who with their thought provoking views, veracity and whole hearted co-operation helped me in doing this thesis.



Vaishali

801532057

ME - CSE

Abstract

In the era of internet, Use of MasterCard is the best way to pay bills. MasterCard payment is increasing on a large scale day by day. MasterCard allows the convenience of spending on credit to the owner of owner of card. This means that the lack of cash availability at that time is not concern for a MasterCard holder since he/she can spend and purchase on credit and pay conveniently at a later date. Before giving a credit loan to borrowers, bank decides who is bad (defaulter) or good (Non-defaulter) borrower. The prediction of borrower status i.e. in future borrower will be defaulter or non-defaulter is a challenging task for bank. The defaulter prediction is a binary classification problem. There are various existing algorithms for checking the default of MasterCard customer such as Support vector machine, Decision tree, Random forest, linear model etc. As we know defaults of MasterCard customer is increasing day by day so current methods will not be proficient in future, hence there is a need to join two or more methods to improve the performance of present models using ensemble methods. Thus various machine learning models have been explored and analyzed. The models have been evaluated based on various performance metrics. At last models have been compared with existing models to evaluate the accuracy.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CNP	Card Not Present Exchange
EMV	Europay MasterCard Visa
BPA	Bayesian Parameter Averaging
BMA	Bayesian Model Averaging
GLM	Generalized Linear Models
MR	Multinomial Regression
SVM	Support Vector Machine
DA	Discriminant Analysis
FDA	Flexible Discriminant Analysis
MDA	Mixture Discriminant Analysis
PLS	Partial Least Square
MARS	Multivariate Adaptive Regression Spline
SMOTE	Synthetic Minority Over-Sampling Technique
AUC	Area Under Curve
PCI	Payment Card Industry
DSS	Data Security Standard
SSC	Security Standards Council
CSV	Comma Separated Values
PIN	Personal Identification Number

Chapter-1
Introduction

The usage of internet is growing exponentially on large scale. So, the use of online shopping and online payment is also growing day by day. There are several procedures proposed for default of MasterCard such as genetic algorithm, and support vector machine etc. The use of internet data is growing exponentially, so the distinct procedure will not be effective. Therefore to overcome this difficulty one procedure can be joined with another procedure, thus it can work extra powerfully as compared to distinct one. The mixture of the different procedures into the only one procedure is known as ensemble model.

1.1 What is MasterCard?

MasterCard is a sort of money related record. It is an installment card issued to clients (cardholders), By utilizing MasterCard, clients can offers a bank's cash rather than their own to pay for an item or administration today, and after some time, they reimburse the bank. For the advantage of utilizing another person's cash, clients will frequently need to pay enthusiasm, obviously with different sorts of advances [1].

1.2 Usage of MasterCard

A MasterCard issuing organization, for example, a panel or MasterCard union, goes into concurrences through traders so as to acknowledge their MasterCard. Traders frequently advertise which passes they acknowledge via showing acknowledgment symbols – for the most part gotten since logos – or this might be imparted in signage either in foundation or organization material.

The charge card guarantor issues a MasterCard to a customer at the same stage or later a record has been affirmed thru the credit supplier, which require not to be indistinguishable element from the card backer. The cardholders would then be able to use it to make purchase at traders tolerating that card. At the time when purchase is done,

the card guarantor consents to pay the cardholder. The cardholder demonstrates agree to fee by designing a receipt with a record of the card points of interest and showing the add up to be funded or by arriving an individual distinguishing proof number. Additionally, numerous vendors now acknowledge verbal approvals by means of phone and electronic approval utilizing the Internet, so called as card not present exchange (CNP).

Various types of usage are:

1.2.1 Grace Period

A MasterCard grace era is that stage when the cardholder needs to fee the adjust previously interest is evaluated based on remarkable adjust. Grace periods might shift, yet for the most part lies between twenty to fifty five days relying upon the sort of charge card and the allotting bank. A few approaches take into account reestablishment after specific conditions are met.

More often than not, on the off chances cardholder is late giving the adjust, fund charges will be figured and the effortlessness time frame did not make a difference. Fund charges acquired rely on upon the effortlessness time frame and adjust; with most Visas there is no elegance period if there is any exceptional adjust since the past proclamation or charging phase.

1.2.2 Interest Charges

Credit card guarantors for the most part postpone interest charges if the balance is ponied up all required funds every month, except normally would charge complete enthusiasm on the whole extraordinary adjust since the time of each purchase if the aggregate adjust is unpaid.

1.2.3 Parties Involved

Various parties are involved in the usage such as insurance providers, acquiring bank, card holder, affinity partner, merchant, transaction network, independent sales organization, card-issuing bank.

1.2.4 MasterCard Register

A MasterCard enlist is an exchange enroll utilized to guarantee the expanding balance payable from utilizing a MasterCard is sufficient beneath as far as possible to bargain for approval installments and holds however not gotten through the bank and to effortlessly look upward past exchanges for bargaining and arranging.

The enlist is an individual record of saving money exchanges utilized for credit card buys as they influence finances in the ledger or the accessible credit. So as to check number and the code section demonstrates the credit card. The left funds after the buys of good are displayed in balance column. At the point when the credit card installment is made the adjust as of now mirrors the assets were spent. In a MasterCard entrance, the store section demonstrates the accessible credit and the installment segment indicates total owed, their whole being equivalent to the card limit.

Each check composed, platinum card exchange, money withdrawal, and credit card charge is entered physically into the paper enroll day by day or a few times for each week [2]. MasterCard enlist likewise alludes to one exchange record for each Visa. For this situation the booklets promptly empower the area of a card's current accessible credit when at least ten cards are being used.

1.3 Types of MasterCards

a) **Business MasterCard:** Business MasterCard are specific charge cards allotted for the sake of an enrolled professional, and commonly they must be utilized for professional purposes. Their utilization has developed in late eras. In 1998, for example, 37% of private companies announced utilizing a professional MasterCard; by 2009, this figure had developed to 64% [3]. Professional charge cards deal various components particular to organizations. They as often as possible offer uncommon rewards in ranges, for example, shipping, office supplies, travel, and business innovation. Most guarantors utilize the candidate's personal rating while assessing these applications.

b) **Secured MasterCard:** A secured MasterCard is a sort of Visa protected by a store account claimed by the cardholder. Regularly, the cardholder must store in the vicinity of 100% and 200% of the aggregate sum of credit craved. Accordingly if the cardholder puts down $1,000, they will be given credit in the scope of $500–1,000. Sometimes, Visa guarantors will offer motivations even on their secured card portfolios. In these cases, the store required might be fundamentally not as much as the required credit constrain, and can be as low as 10% of the coveted credit restrict. This store is held in a unique investment account. Charge card backers offer this since they have seen that misconducts were quite lessened when the client sees something to lose if the adjust is not reimbursed.

c) **Digital MasterCard:** A digital MasterCard is a computerized cloud-facilitated virtual portrayal of some sort of recognizable proof card or installment card, for example, a charge card.

d) **Prepaid MasterCard:** A "prepaid card" is not a genuine MasterCard [4], since card sponsor offered no credit: the cardholder expends cash “deposited” by a former cardholder or another person, for example, a parent or business. In any case, it conveys a Visa mark and can be utilized as a part of comparable routes similarly same as it were a MasterCard [4] unlike debit cards, prepaid MasterCard by and large don't require a PIN. A special case are prepaid MasterCard with an EMV (Enhanced movement vehicle is a specialized standard for savvy installment cards and for installment terminals and mechanized teller machines that can acknowledge them) bit. Such cards need a PIN if installment is prepared by means of Chip and PIN innovation. In the wake of buying the card, the card taker stacks the record thru any measure of cash, up to the prearranged card cutoff and after that utilizing the card to purchase the similar manner as a common MasterCard.

1.4 Advantages and Disadvantages of MasterCard

The primary advantage to the card owner is comfort. Contrasted with checks and debit card, a MasterCard permits little small-period advances to be rapidly prepared to a card owner who require not compute an equilibrium staying earlier each exchange, provided aggregate charges don't surpass the most extreme credit track for the card.

Diverse nations deal distinctive stages of security. Numerous MasterCard suggest rewards and advantages bundles, for example, upgraded item guarantees at no cost, free misfortune/harm scope on new buys, different protection assurances, for instance, rental auto protection, normal bearer mischance security, and travel therapeutic protection.

MasterCard can likewise suggest a steadfastness platform, where every purchase is remunerated thru focuses, which might reclaimed for money or for items. Inquire about have inspected that competition amongst card systems might conceivably make installment compensates excessively liberal, causing higher costs among traders, therefore really affecting social welfare and its appropriation, a circumstance possibly justifying open strategy interventions [5].

At present, there are MasterCard with 0% introduction APR on no late payments and Balance Transmissions.

1.4.1 Harms to Card Owner

a) **Bankruptcy and high interest:** Small basic *MasterCard* charges are restricted to a settled period, for the most part in the vicinity of six and twelve months, after this period a greater amount is charged. As all *MasterCard* expenses and premium, a few clients turn out to be so obligated to their *MasterCard* supplier that they are headed to liquidation. Some *MasterCard* regularly require a amount of 20 to 30 % after an installment is neglected [6]. In different circumstances, a settled charge is demanded without any modification to loan cost. Now and again all inclusive defaulting may apply: the great defaulting rate is connected to the card on favorable terms by omitting an installment on a disconnected record from a similar supplier. This would lead be able to a snowball impact in which the

customer is suffocated by suddenly high loan costs. Further, most card holder understandings empower the guarantor to self-assertively raise the loan cost for any reason they see fit.

Inquire about demonstrates that a generous division of shoppers (around 40%) pick an imperfect charge card understanding, with some acquiring many dollars of avoidable premium expenses.

b) **Declines self-regulation:** A few reviews have demonstrated that buyers are probably going to spend more cash when they pay with MasterCard. Analysts recommend that when individuals pay utilizing MasterCard, they don't encounter the unique agony of payment [7]. Furthermore, specialists have discovered that utilizing MasterCard can increase utilization of undesirable food [8].

1.4.2 Harms to Society

a) **Overestimated rating for all buyers:** Dealers that acknowledge MasterCard must pay exchange expenses and markdown charges on all MasterCard transactions [9] [10]. Sometimes traders are banned by their credit understandings from passing these charges straightforwardly to MasterCard clients, or from setting a base exchange sum (never again denied in the US, UK or Australia). The outcome is that vendors stay prompted to charge all clients (counting the individuals who don't utilize Visas) higher costs to cover the charges on MasterCard transactions [10]. The incitement can be solid in light of the fact that the shipper's charge is a rate of the deal value, which disproportionate affects the productivity of organizations that have dominatingly MasterCard exchanges, unless adjusted at by raising costs by and large. In US in 2008 MasterCard organizations gathered a normal sum of around $427 each family, along a normal expense amount of around 2% for every deal [10].

1.4.3 Profit to Traders

For traders, a MasterCard exchange is frequently extra protected than different types of installment, for example, checks, in light of the fact that the allotting bank resolves to fee the dealer the minute exchange is approved, paying little respect to whether the customer defaults on the MasterCard installment (aside from honest to goodness debate, which are examined beneath, and can bring about charges back to the shipper). As a rule, cards are considerably more secure than money, since they debilitate robbery by the shipper's workers and lessen the measure of money on the premises. At long last, MasterCard decrease the bank office cost for preparing checks/money and transferring all of them to bank.

Preceding MasterCard, every trader needed to assess every client's record of loan repayment before developing credit. Undertaking is presently implemented by the banks who accept credit hazard. MasterCard can likewise help in safeguarding a deal, particularly if the client don’t have plenty money on her/his individual or financial records. Additional revenue is created in such a way that client can buy products as well as administrations instantly and is less restrained by the measure of trade out his or her pocket and the quick condition of his or her bank adjust. A lot of traders' advertising depends on such instantaneousness.

For every buy, bank charges the shipper a charge (markdown expense) for such administration and there might be sure postponement before concurred installment is gotten by the trader

1.4.4 Cost to Traders

Traders are charged a few expenses for tolerating MasterCard. The shipper is typically charged a contract of about 1 to 4 % of the estimation of every exchange rewarded for by MasterCard [11]. Trader may likewise pay a flexible charge, known as trader markdown amount, for every transaction [9]. In a few occurrences of small-esteem exchanges, utilization of MasterCard might essentially decrease the net revenue or make trader miss

cash on the exchange. Trader with small normal exchange costs or great normal exchange costs are extra disinclined to tolerating MasterCard. Sometimes Trader may charge clients a "MasterCard supplement" (or additional charge), either a settled sum or a rate, for installment by MasterCard. This exercise was denied by maximum MasterCard contracts in the US until 2013, while a noteworthy clearance amongst traders and Visa organizations enabled dealers to demand extra charges. Most retailers have not begun utilizing MasterCard extra charges, in any case, because of a paranoid fear of misplacing clients [12].

Traders in US have been battling whatever they contemplate to be unjustifiably great expenses charges with MasterCard organizations in a progression of claims that begun in 2005. Shippers charged that the two principle Visa preparing organizations, MasterCard and Visa, utilized their imposing business model energy to require exorbitant expenses in a legal claim including the Nationwide Marketing Alliance and real sellers.,

Traders remain additionally necessary to rent or buy preparing hardware, at times this gear is without given of charge by the processor. Shippers should likewise fulfill information security consistence measures which are very specialized and confounded. Much of the time, there is a postponement of a few days before assets are kept into a trader's ledger. Since charge card expense structures are extremely confounded, small traders are off guard to dissect and anticipate charges.

At long last, traders expect the danger of charge backs by buyers.

1.5 Security

MasterCard security depends on bodily safety of plastic card and also the protection of the MasterCard figure. Accordingly, at whatever point a man other than the card proprietor approaches the card or its number, security is possibly traded off. When, traders would regularly acknowledge MasterCard figures without extra confirmation for mail arrange buys. It's currently normal exercise to just boat to affirmed talks as a safety effort to limit fake buys. A few dealers will acknowledge a Visa figure for in-store buys,

so access to the figure permits simple extortion, yet several require the card itself to be available, and want a mark. A stolen or lost card can be crossed out, and this is completed rapidly, would enormously confine the misrepresentation that happen thusly. European banks require a card owner's safety PIN to be arrived for face to face buys with the card.

The objective of the Visa organizations is not to dispense with misrepresentation, but rather to "diminish it to sensible levels" [13]. This suggests extortion avoidance methods would be utilized just if their price are minor than the possible additions from misrepresentation lessening, though high-cost and low-return measures won't be utilized.

For the security purpose of the card three upgrades have been offered with more typical MasterCard system, yet none has demonstrated to help decrease visa misrepresentation until this point. In the first place, the MasterCard by their own are supplanted with comparable seeing alter safe shrwed cards which are proposed to mark falsification extra troublesome. The greater part of MasterCard agree to EMV (Europay MasterCard Visa) standard. Secondly, an extra three or four digit code for card security is currently present on the other side of most of the MasterCard. Partners at all stages in electric installments have perceived the necessity to create predictable worldwide models for safety that record and coordinate both present and rising safety innovations. They had started to address these requirements through associations, for example, DSS, PCI and Secure POS Vendor Alliance.

1.6 Costs

MasterCard presenters (banks) have numerous forms of expenses:

1.6.1 Operating costs

As we all know that there is some cost of administration of a MasterCard portfolio, including all the cost such as from paying the official who undertake the organization to production of plastics, to mailing the announcements, to administrating the PCs that monitor each card owner adjust, to taking telephone calls which cardholder do to their sponsor, to shield clients from extortion rings. Contingent upon the sponsor, promoting projects are additionally a critical segment of costs.

1.6.2 Interest payments

Banks for greatest part acquire the money they by then credit to their customers As they get little premium advances from various firms they may get as far as their customers require while crediting their income to various borrowers at great rates In case the card underwriter charges 15% on money advanced to customers and it charges 5% to acquire the money to advance and the alter sits with the card owner for a year sponsor rise 10% on credit gain. This 10% qualification is the "net interest spread" and 5% is "interest cost"

1.6.3 Charge offs

At the point when card owner turns out to be seriously reprobate on an obligation (frequently at the purpose of six months without installment), the loan boss may announce the obligation to be a charge-off. It will then be recorded all things considered on the indebted person's credit agency reports.

A charge-off is thought to "composed off as uncollectable". To panels, awful obligations and misrepresentation are piece of charge of working together.

Though, obligation is still legitimately substantial, and bank can endeavor together everything for eras which is usually a time period of three to seven years allowed under state law, which is normally three to seven years. This incorporates contacts from inner accumulations staff, or probably, an outside gathering organization. On the off chance that the sum is extensive (for the most part over $1,500–2,000), there is the likelihood of a claim or intervention.

1.6.4 Rewards

Many MasterCard clients get prizes, for example, long standing customer focuses, gift authentications, or money back as an impetus to utilize the card. Prizes are for the most part fixing to obtaining a thing or administration on the card, which could conceivably incorporate adjust exchanges, loans, or other exceptional employments. Depending upon the sort of card, prizes will generally cost the guarantor in the vicinity of 0.25% and 2.0% of the spread. Systems, for example, MasterCard had expanded their charges to enable

guarantors to finance their prizes framework. A few issuers debilitate recovery by compelling the cardholder to call client benefit for prizes. On their adjusting site, recovering honors is normally a component that is exceptionally well covered up by issuers. With broke and aggressive condition, prizes focuses censored drastically into a backer's primary concern, and rewards focuses and related motivators must be deliberately figured out how to guarantee a gainful portfolio.

1.6.5 Fraud

In comparative amounts the qualities vanished in bank card misrepresentation are slight, figured in 2006 cost 7 pennies for every 100 dollars' value of exchanges (seven premise points) [14]. In 2004, in UK, the charge of extortion was above £500 million. When stolen the card, or an unapproved copy is made, most card backers would discount a few or the greater part of charges that client has gotten for things that they didn't purchase. Such discounts would, sometimes, be to detriment of dealer, particularly in case of arranging mails where shipper can't guarantee sight of the card. In a few nations, dealers would lose cash if no ID card was requested, in this way vendors more often than not require ID card in these nations. Charge card organizations by and large certification the dealer will be paid on real exchanges paying little heed to whether the buyer pays their Visa charge. Most keeping money administrations have their individual MasterCard benefits that grip misrepresentation circumstances and screen for any conceivable endeavor at extortion. Workers that are spent significant time in doing extortion observing and examination are regularly put in fraud and Authorization, Risk Management, or Cards and Unsecured Business. Misrepresentation observing accentuates limiting extortion misfortunes while making an endeavor to find those dependable and contain the circumstance. MasterCard extortion is a noteworthy cubicle wrongdoing that has been nearby for a long time, even with approach of chip dependent card (EMV) that was incorporated in a few nations to counteract circumstances, for example, such. Indeed, even with usage of such methods, MasterCard extortion keeps on being an issue.

1.7 Revenues

Equalizing the expenses are succeeding revenues:

1.7.1 Interchange Fee

In adding to charges funded by the card owner, traders should likewise pay trade expenses to the card-issuing bank and the card association. For a run of the mill MasterCard issuer, exchange expense incomes may speak to about a fourth of aggregate revenues [15].

These expenses are commonly from 1-6 % of every deal, except will fluctuate not just from trader to trader (vast trader can arrange bring down rates[15]), additionally from card to card, with business cards and rewards cards for the most part costing the vendors more to process. The exchange expense that applies to a specific exchange is additionally influenced by numerous different factors including: the kind of trader, the trader aggregate card deals volume, the trader's normal exchange sum, regardless of whether the MasterCard were bodily existent, how the data that is necessary for exchange was gotten, particular sort of card, when exchanges were settled, approved and settled exchange sums. Sometimes, shippers add an extra charge to the MasterCard to cover the trade expense, urging their clients to rather utilize money, platinum cards, and even checks.

1.7.2 Interest on Due Balances

Interest differ broadly from card sponsor to card sponsor. Frequently, there are "secret" charges in actuality for introductory phase-frames (as small as 0% for, say, six months), though consistent charges may be as high as 40 %. In U.S. no government constrain is there on the premium or late expenses MasterCard backers may charge; financing costs are fixed by states, for example, South Dakota, have no roof on loan fees and charges, welcoming a few banks to set up their MasterCard operations. Different countries, for instance Delaware, having exceptionally powerless loaning laws. The secret amount never again relates if client doesn’t wage their notices on time, and supplanted by a punishment financing cost (for instance, 23.99 percent) that relates retroactively.

1.7.3 Expenses Accused to Clients

Real charges are for:

a) Charges that bring about surpassing as far as possible on the card (regardless of whether purposely or by oversight), brought over breaking point expenses.
b) Returned check charges or installment handling expenses (e.g. telephone installment charge).
c) Late or late installments.
d) Loans and accommodation checks (regularly 3% of the sum).
e) Exchanges in a remote money (as much as 3% of the sum). A couple of budgetary foundations don't charge an expense for this.
f) Enrollment expenses (yearly or month to month), now and again a rate of as far as possible.
g) Conversion scale loading charges (once in a while these won't not be accounted for on the client's announcement, notwithstanding when applied). The variety of trade rates connected by various MasterCard can be extremely significant, as far as 10% as indicated by a Lonely Planet report in 2009.

1.8 What is Default of MasterCard?

Credit card default happens when you've turned out to be seriously reprobate on your credit card installment. It's a genuine charge card status that not just influences your remaining with that credit card guarantor, additionally your credit remaining when all is said in done and your capacity to get affirmed for credit card, advances, and other credit-based administrations.

At the point when installments are not set aside a few minutes and as indicated by the assention marked by the card holder, the record is said to be in default. When you acknowledge a MasterCard, you consent to specific terms.

For instance, you consent to make your base installment by the due date recorded on your financial record. In the event that you miss the base MasterCard installment six months

consecutively, your MasterCard will be in default. Your MasterCard guarantor will probably close your record and report the default to the credit departments.

In the months paving the way to charge card default, your (late) installment status will be accounted to the three noteworthy credit agencies and your credit rating will be affected by the delay of your installments. If you apply for any new MasterCard or advances after a charge card default, your application will probably be denied in light of the fact that banks believe you're at danger of defaulting on any new credit commitments. Actually, a few moneylenders won't favor you at all until you've cleared up the default adjust (or it drops off your credit report).

When your credit is defaulted, you've likely collected many dollars in expenses and premium charges.

Sadly, your alternatives for clearing up the charge card default might be constrained in light of the quantity of installments you've missed for you. Had you reached your credit card guarantor sooner, you may have possessed the capacity to work out a course of action to make installments on the past due adjust and bring your record over into great standing. Now, your charge card guarantor will anticipate that the record will be ponied up all required funds.

Here are your alternatives for managing MasterCard default:

a) **Pony up all required funds**, on the off chance that you have the cash, obviously. To start with, have a go at arranging a compensation for erase where the charge card guarantor expels the record from your credit report in return for installment. A few loan bosses may concur, others won't, however you won't know whether you don't inquire.
b) **Settle the record** for not as much as the sum due. Settling the obligation is likewise a transaction. The loan boss doesn't need to acknowledge a sum lower than the money owed, yet some can be influenced.
c) **File Bankruptcy**. Contingent upon the degree of the default and some other obligations you have, you may consider file bankruptcy to either rebuild your obligation and make it more reasonable or to have it released. Take note of that

bankruptcy remains on your credit report for seven-ten years, so it's not a choice to enter daintily.

d) **Do nothing**. You can disregard the record, maybe choose what to do about it later on not far off. Take note of that the leaser can in any case seek after you for the obligation, show it on your credit report, and may even sue you the length of the statute of constraints is as a result.

e) Before giving a credit advance to borrowers, bank chooses who terrible or great borrower is. Forecast of defaulter position i.e. in future defaulter will be defaulter or non-defaulter is a testing assignment for bank. Credit defaulter forecast is a twofold characterization issue.

f) The issue is to arrange borrower as defaulter or non-defaulter. It is ordinarily wanted for banks to group borrower precisely to deal with their advance hazard better and increment business. However, growing such a model is an exceptionally difficult because of developing interest for advances.

1.9 Structure Of Thesis

The thesis is structured as follows:

Chapter 2: This section talk about the types of machine learning, data mining, ensemble model n its type, R programming language and lastly about the dataset used.

Chapter 3: This section talk about the problem formulation and also defines the objectives of this research work.

Chapter 4: This section talk about the implementation details along with the results of implementation of dataset from various existing algorithms. This section also discusses comparative examination of the technique proposed.

Chapter 5: This section talk about the conclusion and future opportunity of the work.

Chapter-2

Literature Review

This chapter defines the fundamentals of machine learning, its upbringing and varieties of machine learning methods. This chapter likewise examine ensemble methods and its uses. This chapter also defines the core dataset used for the present work.

2.1 Background

Machine learning is a subdivision of software engineering which fundamentally manages the acknowledgment, classification, and learning. It is valid in all parts of computerized reasoning. Machine learning is a branch of software engineering which fundamentally manages the acknowledgment, classification, and learning. It is appropriate in every aspect of computerized reasoning. Machine learning was first created in 1950. It is a branch of manmade brainpower. It is essentially utilized for concentrate enormous information. It is the learning of how the PC understands the actions of the person [1]. Machine Learning is the investigation of how PC reproduction or acknowledge human learning action, to get new information or abilities, to shape the learning structure of the current, to always enhance their execution. It is the center of manmade brainpower, and its application utilized as a part of all ranges of manmade brainpower; it basically utilizes inductive, extensive methodologies [2].

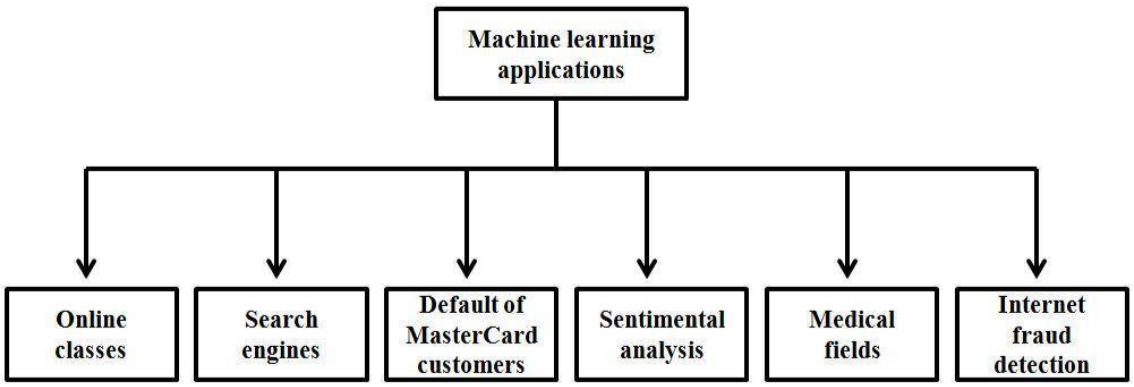


Figure 2.1 Applications of Machine Learning

Figure 2.1 describes the numerous applications of machine learning. One main application of machine learning is default of MasterCard.

2.2 Types of Machine Learning

Machine learning is distributed into four classes namely:

a) Supervised Learning
b) Semi-supervised Learning
c) Unsupervised Learning
d) Re-inforcement Learning

2.2.1 Supervised Learning

Supervised learning utilizes some particular guidelines to predict the yield of the given capacity. In supervised learning, the principle point is to predict the objective components with the help of test component. Supervised machine learning is well-known to be supervised for a purpose and that purpose being machine is delivered with the simple data from the customer end. Documents could be in different methods such as MATLAB file, csv (comma separated values) sheet, excel sheet or some other format. Thought behind supervised learning being device will acquire the circumstances from the records input and do forecasts centered on that. Naïve Bayes technique is the finest example for presenting in what manner this work is finished. Concerned candidates can read Naïve bayes execution from a book entitled “Data Smart” printed by a data researcher himself and fully the thoughts and ideas are executed easily in Microsoft Excel. Supervised Learning is further distributed into two types:

a) Classification

Classification is the approach utilized for classifying the information. In this approach, target esteem is set either non default (1) or default (0). This kind of approach fundamentally utilized when classifying methodology is utilized as a part of the issue i.e. sickness is harmful or not, email is spam or not, mastercard

installment is default or not. There are different applications for classification calculation, for example, discourse acknowledgment, penmanship acknowledgment, biometric, design ID, arrangement of the archive, web crawler and so on. Most normal and important case of classification is to check whether the installment is default or not-default.

List of algorithms supported by classification approach:

i. Naïve Bayes Classifier
ii. Flexible discriminant analysis
iii. Decision Tree
iv. Neural network
v. K-nearest neighbor
vi. Support Vector Machine

b) Regression

Regression is quite inverse of classification. In arrangement approach, objective variable are set either genuine or false while in regression objective variable is set into genuine value. In regression, the yield variable is consistent or genuine value. Regression is additionally separated into ten sections i.e. straight regression, ridge regression, logistic regression, LAD regression, lasso regression, quantile regression, jack-knife regression, Bayesian regression, regression in unusual space, ecological regression and logic regression.

List of algorithms supported by regression approach:

i. Neural network
ii. Generalized Linear model
iii. Linear regression
iv. Decision Tree
v. Nonlinear regression

2.2.2 Semi-supervised Learning

It is mixture of knowledge of data from the tagged as well as untagged data. It basically utilizes the quality of both the learning techniques, i.e. unsupervised as well as supervised. Application of semi-supervised learning is face recognition.

2.2.3 Unsupervised Learning

Unsupervised learning is totally inverse of supervised learning. There is no requirement of output in unsupervised learning. Unsupervised learning mainly agreements with grouping of data. Applications of unsupervised learning are bio-informatics, compression of image, association analysis etc. Unsupervised learning, as the term suggests is interrelated to machine learning by means of itself through some procedures without interfering of the consumer giving the response of data. Most general unsupervised machine learning is Clustering. In clustering method the main role is of distance formula in determining the mean of the points falling below one area.

2.2.4 Re-inforcement Learning

Reinforcement Learning is a kind of Machine Learning, and additionally a branch of Artificial Intelligence. It enables machines and programming specialists to consequently decide the ideal behavior inside a specific context, keeping in mind the end goal to expand its execution. Straightforward reward input is required for the specialist to learn its behavior; this is known as the reinforcement flag.

2.3 Ensemble Model

Ensemble model is a mixture of at least two machine learning model (procedure).Ensemble model is more productive than individual models. In machine learning, group model is consolidating at least two models to improve prediction, accuracy and robustness as related to individual model distinctly. While performing, group model in the beginning put training dataset into various models after that, select the finest model suitable for the dataset. In this work, the study is done by consuming six machine learning approaches i.e. Accuracy, Confusion matrix, Receiver operating

characteristics (ROC) curve, Sensitivity, Specificity and Kappa value. After that execution of k fold validation is completed on top five models.

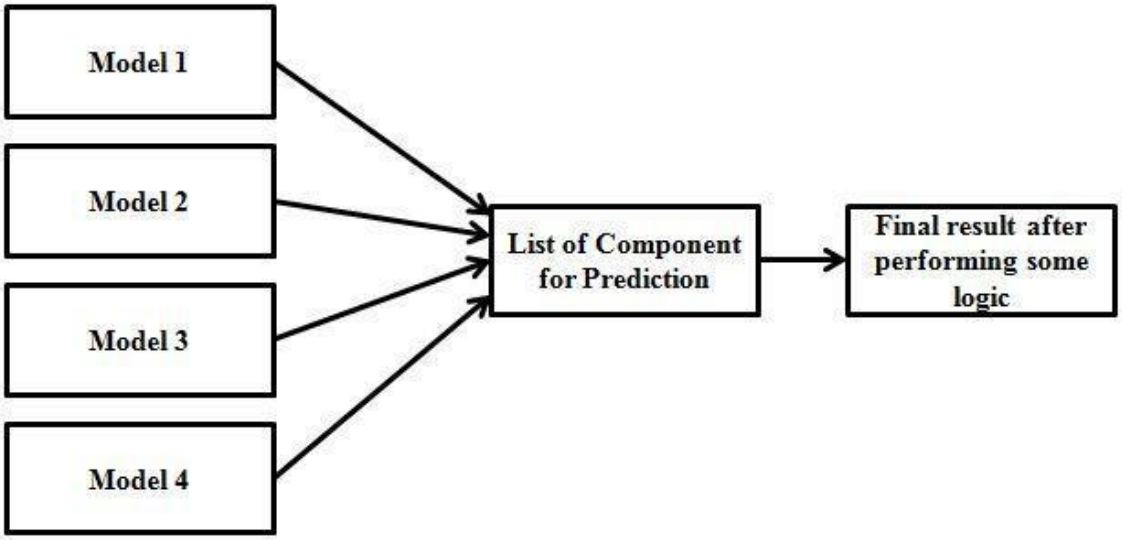


Figure 2.2 Process of Ensemble model

Application of ensemble model:

a) Kernel Factory

Kernel factory is fundamentally an ensemble technique on kernel machine. A kernel is a technique utilize in machine learning. It is mostly used for shape recognitions. In kernel method, there is a job called kernel job. Kernel is fundamentally a resemblance function which is utilized for determining, that how related the two machine learning procedure are. For example, suppose there is a job for text arrangement. This can be finished by two techniques, firstly proceed with written data as training data after that, compute feature and place those features in machine learning procedures and calculate the performance of separate text and after that calculate the resemblance between each written data, second technique is kernel method. In this technique, define kernel task and with respect to that kernel task calculate the resemblance between two written data. In this work there is one kernel related procedure that is kernel support vector machine.

b) Improving algorithm performance

Ensemble model not only associate procedure but similarly rises the performance of the procedure. Suppose naïve Bayes procedure is applied on some dataset. Suppose accuracy for individual dataset is very low, so accuracy can be improved by two technique. First one is to utilize another procedure and the second one, associate this procedure with other procedures for increasing the performance. In this work, second technique is used for increasing the performance of procedures. This application is likewise very useful to improve the effectiveness and the accuracy of the algorithms.

2.3.1 Types of Ensemble Model

There are different types of ensemble model, they are as follows:

a) Bootstrap aggregating (bagging)
b) Bayes optimal classifier
c) Bayesian parameter averaging
d) Boosting
e) Bucket of models
f) Bayesian model combination
g) Stacking

2.3.1.1 Bootstrap Aggregating

Bootstrap aggregating, or bagging, includes having every model in the collection is vote with equivalent weight. Using a casually taken subgroup of training group bagging trains every model in the collection so as to endorse modal variance. As an example, to attain a very high classifier correctness the random forest method joins bagging with random decision tree technique. Although it is commonly applied to decision tree technique, it can be very helpful and can be used for several technique. Bagging is a right superior case of model averaging method.

2.3.1.2 Bayes Optimal Classifier

Bayes Optimum Classifier is only an association technique. In the hypothesis space, is a collection gathering of all the hypothesis. Overall, none of the other method would be able to beat it. Each hypothesis is set with a token proportionate to the likelihood that the showing dataset will be tried from an association if those hypothesis were valid. The token of each hypothesis is increased by the past likelihood of that same hypothesis, to encourage showing records of limited size. Unfortunately, the Bayes Optimum Classifier couldn't be for all hypothesis and purposes executed for any theory however the best straightforward of challenges.

2.3.1.3 Bayesian Parameter Averaging

Bayesian parameter averaging (or BPA), an outfit procedure that attempts to approximate the Bayes Optimal Classifier by reviewing theories from the hypothesis space and merging them using Bayes' law. Not at all like the Bayes perfect classifier Bayesian model averaging (or BMA) can be in every way that really matters realized theories are routinely inspected using a Monte Carlo looking at framework for instance MCMC. For example, Gibbs examining may be used to draw hypothesis that are illustrative of the movement. It has been shown that in particular circumstances when theories are pulled along these lines and found the middle estimation of according to Bayes' law this technique has a typical blunder that is constrained to be at most twofold the ordinary mistake of the Bayes ideal classifier.

Despite the theoretical correctness of this system, earlier work demonstrated trial comes about suggesting that the procedure progressed over fitting and performed all the more awfully in contrast to more direct assembling techniques for instance stowing in any case these conclusions appear from every angle to be established on a misconception of the explanation behind Bayesian model averaging versus show mix. Besides, there have been noteworthy advances on a fundamental level and routine with respect to BMA. Late exhaustive confirmations demonstrate the precision of BMA in factor determination and estimating the high dimensional settings and give exploratory evidence highlighting the part of scarcity upholding priors inside the BMA in decreasing over fitting.

2.3.1.4 Boosting

Boosting includes incrementally building a group via preparing each new model case to underscore the preparation occurrences that past models mis-ordered. Now and again, boosting has been appeared to yield preferred exactness over bagging, however it likewise has a tendency to probably over-fit the preparation information. By a wide margin, the most well-known execution of Boosting is Adaboost, although some more up to date calculations are accounted to accomplish better outcomes.

2.3.1.5 Bucket of Models

A "bucket of models" is a system in which a model decision calculation is used to pick the best performing model for each issue. When used with only a single bucket of models can convey no ideal results over the best model in the set. However when utilized with crosswise over various issues it will commonly make much better results by and large than any model in the set.

The most broadly perceived approach used for model determination is known as cross validation selection. Cross Validation Selection can be summarized as "endeavor them all with the participation set and select the model that performs best". Gating is an approach of Cross Validation Selection. It incorporates the preparation of another learning model to pick which of the models in the bowl is most fitting to deal with the issue. Consistently, a perceptron is used for the gating model It can be used to pick the "best" model or it can be used to give an direct weight to the desires from each model in the bucket.

When a bucket of different models is used with a specific arrangement of issues it may be attractive to abstain from setting up a portion of the models that put aside a long opportunity to prepare. Milestone learning, a meta learning approach that hopes to deal with this issue. It incorporates preparation of only the quick however loose calculations in the basin and after that using the execution of these calculations to help figure out which moderate (yet precise) calculation is well on the way to do best.

2.3.1.6 Stacking

Stacking or sometimes also called as stacked speculation incorporates preparation of a learning calculation to join the desires of a couple of other learning calculations. Firstly the major part of alternate calculations are readied using the available data at that point a combiner calculation is set up to make a last desire using each one of the forecasts alternate calculations as additional sources of information. If a subjective combiner calculation is used at that point stacking can hypothetically communicate with any of the systems depicted in this article despite of fact that a single layer strategic relapse model is regularly used as the combiner.

Stacking consistently yields execution better than any of the already prepared models. Stacking has been successfully used on both the supervised learning backslide portrayal and partition learning and the unsupervised learning estimation. Also, stacking has been used to evaluate bagging's mistake. Stacking has also been observed to perform better than Bayesian averaging model. Top two performers of the Netflix competition utilized blending which may be believed to make a different kind of stacking.

2.4 Data Mining

Machine learning and Data mining are very narrowly interrelated terms. Data mining procedure are the methods that are utilized to get the effects from dataset delivered to the machine. Machine learning is the method which is utilized as a superset though applying records mining methods. Around two hundred record mining methods which could be executed on the dataset are presented in library and the list of all two hundred methods is provided in R programming. However with each data mining method, machine learning class is provided. And centered on that information various data mining methods are applied on the dataset.

Different data mining methods have been used in our work and they are as follows:

a) Decision Tree
b) Random Forest
c) Support Vector Machine
d) Linear Model

e) Discriminant Analysis
f) Bagging
g) Generalized Linear Models (GLM)
h) Partial Least Square
i) Multinomial Regression (MR)
j) Multivariate Adaptive Regression Spline
k) Stochastic Gradient Boosting

2.4.1 Decision Tree

One of the best mutual data mining model is decision tree model. The reason behind the popularity of decision tree is that the resultant model is easy to recognize. The approach used in this procedure is recursive partitioning. Usually, a graph similar to tree of possible outcomes and decisions works as a decision support device for a decision tree. And, a decision tree knowledge utilizes a decision tree as a projecting model observation around a point to decision about the point's goal value. Tree model that can take simply a fixed set of values called as classification tree. The decision trees where continuous values are taken by the target variables is known as regression trees.

In this work three different methods of decision tree are used, they are as follows:

i. **rpart:** The traditional procedure is applied in rpart package. It is equivalent to ID3/C4 and CART.
ii. **ctree:** conditional tree procedure is executed in party package. It construct tree in a restricted inference framework.
iii. **oblique.tree:** Function used in this package is obliqueTree, by mean of binary recursive separating, only linear mixtures of the input and oblique splits.

2.4.2 Random Forest

Random forests are greatly utilized when we have huge training records and mainly a very vast amount of input variables (hundreds or even large amount of input variables). The procedure is effective with respect to a huge amount of variables as it repetitively

subsections the variables presented. To view the comparative importance of every variable make use of the important key.

Using ten or more of decision trees a random forest model is usually made up. On increasing the quantity of trees make use of the Errors key to check the frequency of decline of the model error.

Mainly random forests approach are utilized in classification and regression. They are ensemble learning technique in itself. Finding the type of Regression and classes in the mean forecast of distinct trees. For the benefit of classification level random forest method utilizes multiple decision trees. In this work we are using rf method of random forest.

2.4.3 Support Vector Machine

A Support Vector Machine (SVM) examines for so named support vectors which are record facts that are set up to lie at the verge of an region in universe which is a borderline from one session of points to a new. In the vocabulary of SVM we discourse about the universe between areas containing record points in diverse classes as being the boundary between those classes. The support vectors are utilizes to recognize a hyper plane (when we are speaking about two dimensional record i.e. a line, or about numerous dimensions in the records) that separates the classes.

SVM is divided into two categories namely:

i. Non-Linear SVM (i.e. kernel based)
ii. Linear SVM

2.4.4 Linear Model

A linear regression model is the old-fashioned method for adjusting a numerical model to data. It is suitable when the objective variable is continuous and numeric. We can explained linear regression as the modeling of connection of scalar independent and dependent variable. Linear regression is called as simple linear regression when there is a single independent variable. And if there are more two or more independent variable then this is the case of multivariate or multiple linear regression.

Here the old-fashioned linear regression model is extended to goals with non-gaussian i.e. non-linear distribution. Linear regression prototypes are iteratively suitable to the records after converting the objective variable to a constant numeric.

2.4.5 Discriminant Analysis

Discriminant analysis uses constant variable dimensions on dissimilar sets of items to high spot features that discriminate the sets and to utilize these dimensions to categorize new items. Common usages of the technique have been in biotic classification into types and sub-types, categorizing applications for giving a loan, insurance and MasterCard's into high risk and low risk categories, etc.

On the basis of observed forecaster variables discriminant analysis (DA) is normally used to construct a descriptive model of set discrimination and it is a multivariate numerical technique. The common aim of DA are to examine dissimilarities between sets, to distinguish sets effectively, to find major discriminating variables, to categorize new remarks into pre-existing sets and to carry out postulate analysis on the change between the estimated sets and to categorize new remarks into pre-existing sets. A variable choice method was directed previously to discriminant analysis, in order to decrease the size of data and to mark discrimination models tougher.

In this work we are using two Discriminant Analysis methods namely:

i. **Flexible Discriminant Analysis (or FDA)**

   Flexible discriminant study is a valued instrument for multigroup ordering. With a huge amount of forecaster, one can discover a moderate quantity of discriminant coordinator tasks that are "optimum" for splitting the groups (FDA BY OPTICAL ANALYSIS). It uses the function fda in mda set and the defaulting linear regression technique.

ii. **Mixture Discriminant Analysis (or MDA)**

   A technique for arrangement based on combination models. It is an addition of linear discriminant study. It uses the function mda in mda sets.

2.4.6 Bagging

Bagging is useful in analogy of decision making. The area in which the managers have no knowledge, they seeks experts' for their advice in that area. Advice-givers must accompaniment each other's skill instead of being duplicative. The type of bagging method used in our work is treebag.

i. **treebag:** using caret package

2.4.7 Generalized Linear Models (GLM)

Generalized Linear Model (GLM) permits for plenty of dissimilar, non-linear prototypes to be verified in the perspective of regression. GLM is the mathematical structure used in various statistical studies such as moderation, multiple regression, moderation, analysis of variance and meditation.

GLM is a supervised procedure with a classic arithmetic method (Supports a huge amount of text, transactional data and input variable) GLM is a supervised procedure for regression and/or classification. GLM implement linear regression method for uninterrupted target and logistic regression method for classifying binary targets.

2.4.8 Partial Least Square

Partial least square (or PLS) regresssion is a technique which simplifies and associate structures from multiple regression and principal component analysis (or PCA). It is essentially advantageous when we need to figure a collection of needy factors from a major collection of free factors (i.e., predictors).

PLS is useful in finding the important relations among two mediums (X and Y), i.e. a latent variable methodology to molding the covariance arrangements in these two places. A PLS prototype will attempt to find multidimensional route in the X places that describes the extreme multidimensional variance route in the Y place. PLS regression is mainly suitable when among X values there is multi collinearity and the mediums of forecasters has added variables than opinions.

2.4.9 Multinomial Regression (MR)

When the reliant variable is dual (binary) then logistic regression is suitable regression to perform. This type of regression is a predictive analysis just like all other regression analysis. Logistic regression is useful in explaining the relationship between one reliant dual variable and other interval or ratio-level, nominal, ordinal independent variable.

At times logistic regressions are hard to interpret; the Statistics tool simply permits you to conduct the study, and in simple English understands the output. Data for which the first-order calculation is adequate with linear relationship, for such data linear regression is very helpful. Also linear regression is not optimal or appropriate for plenty of applications. Therefore, an alternative regression method well suited for this type of data is logistic regression (LR).

2.4.10 Multivariate Adaptive Regression Spline

Multiple adaptive regression spline (MARS) is a regression method that models the connection among target response and numerous predictor variables. The main power of MARS is its capability to identify and the interpretability of difficult interactions between an objective variable and a group of forecaster variables. The MARS model is characterized by moderate linear functions which associate interactively and additively.

The MARS model is planned in such a manner that can forecast continuous mathematical outcomes like a mobile phone monthly average bill of a customer. Mars is also skilled of generating high value probability prototype for a no/yes output. Mars also performs interaction detection, variable transformation, self-testing, and variable selection, all robotically and at great speed.

2.4.11 Stochastic Gradient Boosting

Gradient boosting is method for classification and regression problems, which yields a guess model in the form of a collaborative of weak forecast models, typically conclusion trees. It constructs the model in a phase-wise manner like former boosting approaches do, and it simplifies them by permitting optimization of a random differentiable damage function.

Gradient boosting builds additive regression prototypes by consecutively fitting a modest parameterized task (base learner) to present "pseudo"-residuals using least squares at every repetition. It is displayed that gradient boosting speed of both the estimate execution and exactness can be significantly enhanced by including randomization into the process. Specifically, at every repetition a subsample of the instruction data is drawn at unplanned (without replacement) from the complete training data group. This randomly nominated subsample is formerly used in place of the complete model to adjust the base beginner and calculate the model modernize for the present iteration. This randomized methodology also growths strength against overloading of the base beginner.

2.5 R Programming Language

R is a free (open source) software, basically used for doing statistics calculation. R is not only utilized for statistics calculation but it is also a good programming language (like C or JAVA). The code we write in R is too much alike to python. R is full of libraries and also a GNU project. Around five thousand packages are present in R. The advantages of having such a huge number of packages is that it saves time and avoid a countless arrangement of line of code. Because of the simplicity of R, it is thought to be the finest platform. R also shows graphical calculation very remarkably as related to Matlab and Mathematic. We can use R in whichever operating system as R is platform independent. Languages like JAVA, C++ etc. can be integrated with R. R can be used for finding missing terms. R is also useful in business. Standard companies that are using R, are Facebook (Status analysis, colleague and friends interaction status), Google (in effective creation of online add), Bing (helpful in social search as it increases the awareness).

With respect to machine learning R is thought to be the finest programming language as compared to MATLAB, SAS, SQL, JAVA, Python, C++, and Mathematic. Although other languages like Julia which is considered to be more efficient than R, but still R is considered to be the finest programming language for graphical and statistical calculation and fittest platform for executing machine learning procedures.

R is available in many different interface. These interfaces are:

a) **R studio:** It is a coordinated advancement condition. For running R studio initially introduce R. Most recent variant of R is 3.2.1. This is steady form which was out on 18 June 2015.

b) **Rattle:** It is likewise a graphical UI. It is essentially utilized for information mining (include choice. Include extraction and so forth.)

c) **R commander:** It is additionally a graphical UI. Stable adaptation of R administrator is 2.0.0, out on 21 August 2013.

d) **RKWard:** This is the expanded graphical UI and coordinated improvement condition for R. It is fundamentally composed in C++ and ECMA script. Stable adaptation of this product is 0.6.3, out on 7 march 2015.

2.6 Dataset and Description of its Features

The data set for default of MasterCard is taken from UCI repository. There are total 30,000 instances and twenty four features [2]. This dataset is imbalanced with 0(23364) and 1(6636). In order to balance this imbalanced data, the balancing algorithms such as ROSE, SMOTE, Neighborhood Cleaning Rule, Down-Sampling, Up-Sampling, Tomek Link, one Side Selection and Edited Nearest Neighbor are explored. Out of these algorithms, SMOTE is chosen for balancing the data. Since some of the considered features may have higher importance than others in predicting the default and non-default, Gini and correlation technique are used to determine the feature importance. Feature selection is a process of keeping only those features which contribute in the prediction in a majority manner. Gini index gives the percentage of contribution corresponding to each feature. So, after doing feature extraction (by applying Gini, correlation) we are left with fifteen important features.

This work utilized a binary target variable, named as default installment (Yes = 1, No = 0). The presented work reviewed literature and utilized the below mentioned twenty three factors as the informative factors:

X1: Total credit given. This amount includes the individual and his/her family credit.
X2: Gender of customer, where: 1 = male and 2 = female.

X3: Education of customer, where: 1= graduate school; 2 = university; 3 = high school and 4 = others.

X4: Marital status of customer, where: 1 = married; 2 = single and 3 = others.

X5: Age of customer in years.

X6-X11: Past payment record history from April to September, 2005 that was tracked as follows:

X6 = Repayment status for the month September, 2005;

X7 = Repayment status for the month August, 2005;

X11 = Repayment status for the month April, 2005.

where, repayment status:

1 = Amount duly paid;

1 = payment delayed by 1 month;

2 = payment delayed by for 2 months;

...;

8 = payment delayed by 8 months;

9 = payment delayed by 9 months and above.

X12-X17: Bill amount for various months.

X12 = Bill amount for September, 2005;

X13 = Bill amount for August, 2005;

X17 = Bill amount for April, 2005.

X18-X23: Previous paid amount.

X18 = Amount paid by customer in September, 2005;

X19 = Amount paid by customer in August, 2005;

X23 = Amount paid by customer in April, 2005.

Chapter-3
Research Formulation

This section focused on the problem statement along with research gaps and describes the objectives of this research work.

3.1 Problem Statement

The utilization of web is expanding exponentially step by step. In this way, the instances of default installment are additionally expanding step by step. There are numerous calculations proposed for default installment, for example, multinomial bayes classifier, support vector machine, and genetic algorithm and so on. The utilization of web information is expanding exponentially, so the individual calculation would not be productive. Along these lines, to beat this issue one calculation can be consolidated with another calculation, in this manner it can work all the more productively when contrasted with individual one. The mix of the distinctive calculations into the single calculation is called ensemble model.

Ensemble modeling picks up popularity on the grounds that numerous associations sent PC programming to run such a model. A portion of the product are SAS prescient investigation, IBM prescient examination and so forth. In this way, the utilization of ensemble the model assumes an imperative part today.

Because of expanding instances of default installment in web, productive approach has dependably been the most looked into range. So it is dependably a testing task to propose such a model which works effectively for all sort of information arrangement. R programming dialect is one of the important stages for testing the machine learning models for the given dataset. Machine learning is the measurements approach so the utilization of Microsoft Excel assumes an essential part. Just ensembling the model is insufficient for a dataset, feature selection likewise assumes an essential part while group the model.

3.2 Research Gaps

As many calculations are proposed for default installment. Distinctive calculation works productively for the diverse dataset and diverse information form. For proposing the calculation that suited best for all information arrange ensemble model assumes an imperative part. Although numerous information sorts are the blend of at least two unique information form.

3.3 Objectives

The target utilized for this work are examined below:

a) To analyze the different machine knowledge models on default dataset for various features, partition etc.
b) To play out the relative analysis for various machine learning models on the premises of confusion matrix, specificity, error rate (ER), sensitivity and accuracy.
c) To propose the ensemble model for the dataset on the premises of execution measurements.

3.4 Research Methodology

Machine learning model is executed on R programming linguistic where dataset is default dataset other stage utilized for this reason for existing is Microsoft Excel. Research methodology utilized is:

a) The dataset for default payment is collected from UCI repository.
b) Firstly, models are implemented on raw data. There are twenty four features in the data set.
c) Feature selection is done using correlation and Gini index methods. Gini index is the technique presented in random forest basically used for finest feature selection.
d) Now match the performance of machine models between cleaned data and raw data.
e) Match the performance of records for various division of testing and training data.

f) Analysis of outcome on the base of confusion matrix, error rate, accuracy, specificity, ROC, sensitivity.
g) After examining the result, ensemble the best five models and prepare a graph for those best model for the cleaned dataset.

3.5 Effective Utilization of Rattle Tool and R Programming

R Programming is the standard factual programming which is utilized as a part of machine learning field all through the world. This has its own particular points of interest over different linguistics and Python language is the nearest to what R is prepared to do. R has a library characterized of rundown of information mining calculations accessible for machine learning at goo.gl/3DWT2s. There are more than two hundred models accessible for various sort of informational collections, considering regression, clustering, characterization, multivariate and so on sorts of investigations required.

Rattle is an apparatus like weka, yet Rattle being all the more intense as it is totally worked through R programming only. Rattle is exceptionally useful for learners in field of machine learning. Diagram graphs can be worked in it. Methodology for machine learning and information mining is likewise simple through rattle. Graphical UI is constantly simple to get to and effortlessly justifiable to the learners. But this apparatus has its own restrictions, not more than four-five models can cooperate by depending just on this instrument. Most extreme preferred standpoint of R programming must be taken for getting to up to fifteen-thirty models through R programming utilizing the libraries characterized.

3.6 K-fold Validation

K – Fold is most critical piece of general examination. Explaining it in exceptionally fresh form, this is the idea of checking the consistency of the model while anticipating the objective incentive with the utilization of machine learning calculation. For our situation, we have done 10-fold approval, which means, dataset and diverse information mining and machine learning models will be checked with various partition for ten distinct circumstances, guaranteeing each time distinctive preparing and testing informational

index segments are picked and after that checking the precision for consistency. In the event that the precision is reliable and results are worthy, at that point we reason that the model is in fact a decent model for forecasting the objective value.

3.7 Analysis and Significance of Ensemble Models

Thorough examination of ensemble prototypes will be done in the Result and Implementation section, i.e. section 4.9. Groupings of best five machine learning procedures will be tested so as to select the finest one out of all the permutation.

Ensemble demonstrating is the most critical part of our exploration. Joining at least two machine learning models and afterward contrasting those combinations to locate the most ideal blend, which will be suggested as the most ideal ensemble for issue to be solved.

Chapter-4

Implementation and Results

This section presents executed models on default of MasterCard customer dataset, feature selection of the dataset which comprises selection of important features. This section also defines comparison based on various partition for ROC and Accuracy values. Model contrast and K-fold validation for top five models are also discussed. This section also describe ensemble modeling.

4.1 Implementation Environment

Dataset that we are using in is taken from UCI repository. It is default dataset. Platform that we are using in our work is R programming language and MS excel, R is basically used for executing various machine learning prototypes on the default dataset and MS excel is utilized for matching the performance of various prototypes. The prototypes are established in R. The requirement of machine which is needed for executing purpose are - 64-bit Windows Operating System and Intel(R)Core(TM)i3-2350M CPU @ 2.30GHz 2.30 GHz with 4 GB RAM.

4.2 Proposed Method

The method is defined in Figure 4.1. Firstly, raw experimental data set is downloaded from UCI warehouse [16]. Secondly, Data preprocessing is performed. Imbalanced data is made balanced using SMOTE techniques. After that fifteen machine knowledge models are implemented (Such as Random forest, Ada model, decision tree, SVM, neural network etc.) on balanced data set. For our data set we divided the data into 80-20 partition for training-testing the model. After applying various models, the accuracy is obtained. Removal of irrelevant features is done using correlation and GINI index techniques. In the next phase, cleaned data generated by above steps is used for model building and predictions using those same classification models (refer to table 2) and

keeping the ratio of training and testing same as before (i.e. 80-20). By analyzing the predictions of machine learning algorithms, the top performing models can be found. After model evaluation, selection of top five model is completed w.r.t accuracy and area under the curve i.e. ROC curve, these finest models helps to achieve k-fold validation and next task is to ensemble those best models. Figure-4.1 shows the flow chart of proposed approach.

a) Data Preprocessing

A data set is said to be dirty or imbalanced if there is no approximate equality between classification categories. Frequently real-word data mainly consists of simple examples with a little percentage of engaging example [17]. It is additionally the case that the cost of misclassifying an unusual (interesting) example as a typical illustration is frequently significantly higher than the cost of the turnaround blunder. The execution of machine learning calculations is normally evaluated utilizing predictive accuracy. However, this is not suitable when the information is imbalanced and additionally the expenses of various mistakes change markedly. It gives one-sided expectations and misleading accuracy. This is due to machine learning that calculation doesn't get necessary data about minority class. It is vital to adjust the information so as to maintain a strategic distance from the battle with accuracy as a result of the unequal circulation in dependent variable. There are various machine learning algorithms to clean the data. For our data set we are using Subsampling technique such as SMOTE algorithm to balance the data.

SMOTE: SMOTE stands for Synthetic Minority Over-Sampling Technique [18]. Using K-nearest neighbor and bootstrap this technique falsely output new examples of minority class. The majority category is under-sampled. By randomly fetching samples from majority category population till the minority category becomes a little identify percentage of majority class.

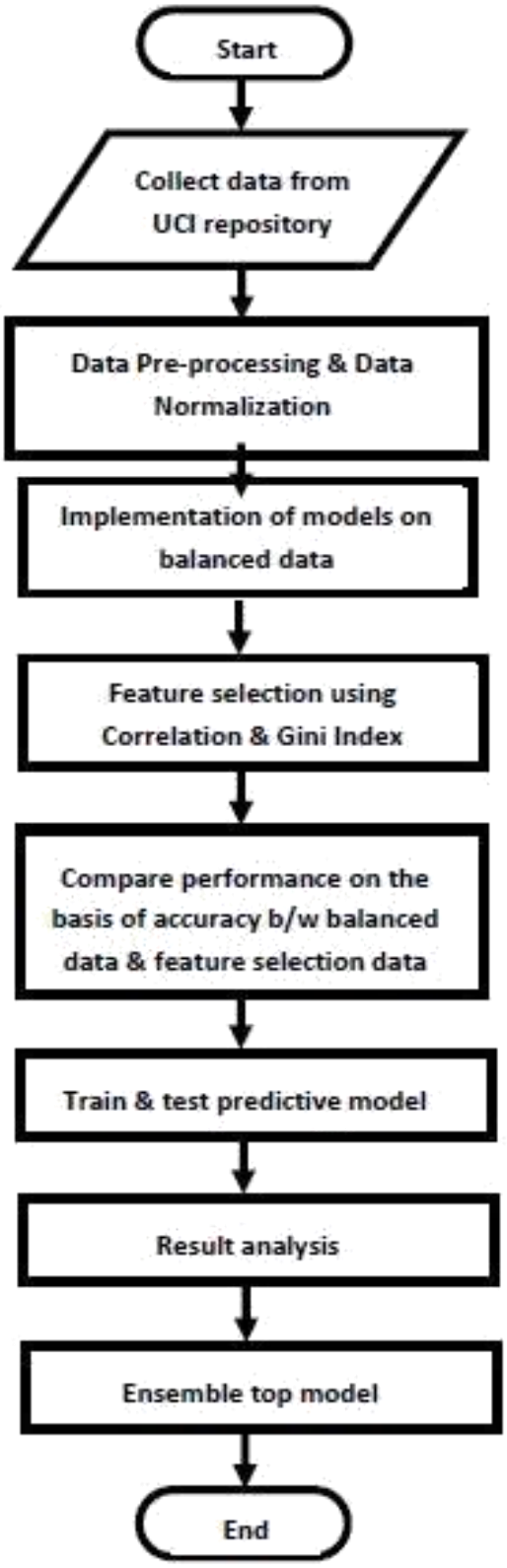


Figure 4.1: Flow chart of Proposed Method

b) Data Discretization

Preprocessing of data is done to reduce the burden of machine learning as original data contain some noise which may be inconsistent or incomplete. To learn various invariants that doesn't make much difference in the symbols meaning but only alter the representation. We use Data Discretization for such alterations. The Data Discretization changes over the information into Discrete information by partitioning the range of a continuous property into interims and lessens the data size. The Data Reduction sub organize remove the superfluous qualities furthermore, decreases the quantity of tuples by examining Here in the data set some column like bill payment have very untidy values, therefore to overcome with these untidy value we do data normalization. In data normalization the data is sorted in increasing order and then divide instances into 10 equal bins. After that rank the first bin as 1, second bin as 2, third bin as 3 and so on up to tenth bin.

c) Feature Selection

Correlation and Gini index technique are used for selecting the important feature in the data set. Correlation is a kind of association or tie-up between two or more object. Gini index is measuring the inconsistency of the distribution. Firstly, we apply correlation technique on data set, on the basis of correlation result removed three irrelevant features that are least important and left with twenty one features. Now after applying correlation we apply Gini index on those twenty one feature data set and found six irrelevant features and after removing those six useless features, we are left with fifteen important features.

d) Machine learning methods

For the prediction of accuracy of default of MasterCard customer data set total fifteen machine learning models are used. The model along with their methods and packages are mention in Table 4.1. To run the models Caret library is used.

Table 4.1: Models used for Analysis

S. No.	Model	Method Argument Value	Package
1	Conditional Inference Tree	Ctree	fastAdaboost
2	Multivariate Adaptive Spline	Earth	Earth
3	Boosted classification tree	Ada	Ada,plyr
4	Stocastic Gradient Boosting	Gbm	gbm,plyr
5	Generalized Linear Modal	Glm	Glm
6	Flexible Discriminant Analysis	Fda	earth,,Mda
7	Penalized Multinominal Regression	Multinom	Nnet
8	Oblique Tree	oblique.tree	Obique.tree
9	Partial Least Square	Pls	Pls
10	Decision Tree	Rpart	Rpart
11	Random Forest	Rf	randomForest
12	Mixture Discriminant Analysis	Mda	Mda
13	Support Vector Machine	Svm	Svm
14	Treebag	Treebag	ipred, plyr,ie1071
15	Linear	Multinominal	Nnet

4.3 Implementation of the Proposed Work

There are numerous procedure existing in machine learning process. These procedure are Stochastic Gradient Boosting, Conditional Inference Tree, Partial Least Square, Oblique Tree, Random Forest, Decision Tree, Generalized Linear Model, Multivariate Adaptive Regression Spline, Linear, Support Vector Machine, Mixture Discriminant Analysis,

Flexible Discriminant analysis, Boosted Classification Tree, Penalized Multinomial Regression. A brief discussion of all the fifteen models used is as follows:

a) **Conditional Inference Tree**: Measurements based approach that utilizations non-parametric tests as part criteria, corrected for different testing to abstain from over fitting. This approach brings about fair indicator determination and does not require pruning.

b) **Multivariate Adaptive Regression Spline**: It is a variation of generalize linear model. It comprises the technique given by Freidman in Multivariate Adaptive Regression Splines [19] and in fast MARS [20].

c) **Boosted classification tree**: This boosting calculation does not require any earlier information about the execution of the week learning calculation [21].

d) **Stochastic Gradient Boosting**: It fabricates the model in a phase wise form like other boosting techniques do, and it sums them up by permitting optimization of a self-assertive differentiable loss work.

e) **Generalized Linear Model**: Generalized Linear Model (GLM) is an adaptable generalization of common direct regression that takes into consideration reaction factors that have mistook appropriation models other than an ordinary distribution.

f) **Flexible Discriminant Analysis**: It is basically utilized for multi-group classifier [22].

g) **Penalized Multinomial Regression (nnet)**: Preparing of neural systems utilizing back-propagation, strong back propagation with or without weight or the altered internationally convergent variant [23].

h) **Oblique Tree**: It is a piece of CART structure. It is utilized when hub is being part to discover slanted hyper plane [24].

i) **Partial Least Square**: It is a statistical technique that bears some connection to principle component regression; rather than discovering hyperplanes of most extreme difference between reaction and autonomous factors, it finds a straight regression show by anticipating the predicted factors.

j) **Decision Tree (Rpart)**: This technique is an augmentation C4.5 grouping calculation [25].

k) **Random Forest (rf)**: Random Forest depends on a forest of trees utilizing irregular inputs. In a random forest, every hub is part utilizing the best among a subset of predictors arbitrarily picked at that hub. In expansion, it is extremely easy to understand as it has just two parameters (the quantity of factors in the arbitrary subset at every hub and the quantity of trees in the forest) [26].

l) **Mixture Discriminant Analysis**: It is a developed variant of direct discriminant analysis. It is basically based on blend of models.

m) **Support Vector Machine**: SVM is an intense technique for general (nonlinear) arrangement and anomalies identification with an instinctive model representation [27].

n) **Linear**: It utilizes direct models to do regression, single stratum investigation of fluctuation and investigation of covariance [28].

4.4 Implementation on raw data

As conversed earlier raw data comprise of twenty-four attributes. Firstly, execution of machine learning prototypes is done on raw record. The judgment is finalized on the basis of various parameters like sensitivity, specificity, ROC curve, accuracy, error rate. The judgment is done on two situations one is data before feature extraction i.e on discretized data and other is after feature extraction. Models that are executed on raw data are decision tree, neural network, random forest, linear model, support vector machine.

a) Accuracy

The accuracy is calculated between the actual value and the predicted value. The Predicted value is predicted output of the respective model. Accuracy is calculated using method

```
If (Actual values == Predicted
        value) Accuracy = 1
else
        Accuracy = 0
```

b) **Error matrix**

Confusion matrix is a name used in classification, also known as error matrix for regression. A 2 x 2 matrix is basically an error matrix categorized by false negative, true negative, true positive and false positive. Lower values in error matrix means the model performance is good.

c) Area Under the Curve

Accuracy of the model is also determined by ROC curve. Area under curve is responsible for determining the ROC curve performance. Model will assume to be perfect iff AUC value is 1. AUC distribution is as shown below:

If AUC lies in range 0.9 to 1.0 than check is said to be brilliant.
If AUC lies in range 0.8 to 0.9 than check is said to be good.
If AUC lies in range 0.7 to 0.8 than check is said to be fair.
If AUC lies in range 0.6 to 0.7 than check is said to be poor.
If AUC lies in range 0.5 to 0.6 than check is failed.

In our work ROC is responsible for the performance of the prototype. Performance is calculated using Area Under Curve (AUC) which is one of the main stuff of ROC.

4.4.1 Executing various Machine Learning Models on raw data

As we have already discussed that the raw data is consist of twenty four features. So firstly we will run our model on the default MasterCard dataset having twenty four attribute. Various parameters like accuracy, sensitivity, error rate, specificity, AUC have been noted down for the performance measurement. Figure 4.2 shows raw dataset with twenty four attributes.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	LIMIT_BAL	SEX	EDUCATIC	MARRIAGI	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT
2	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0
3	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261
4	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549
5	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547
6	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131
7	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024
8	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944
9	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567
10	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719
11	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912
12	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731
13	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668
14	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870
15	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894
16	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512
17	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211
18	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104
19	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599
20	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0
21	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0
22	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930
23	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316

Figure 4.2: Raw dataset

The table below shows the performance of various models on different parameters on raw data.

Table 4.2: Results of proposed method on raw data

Method	Accuracy	AUC	Error Rate	Sensitivity	Specificity
Nnet	72.09	0.733	0.199	0.781	0.755
RF	77.14	0.743	0.209	0.781	0.8122
Oblique	79.89	0.756	0.155	0.743	0.739
Svm	75.72	0.649	0.179	0.795	0.691
Linear	77.18	0.691	0.224	0.791	0.819
Earth	76.01	0.709	0.189	0.808	0.851
Treebag	76.77	0.711	0.172	0.777	0.789

4.4.2 Executing various Models on Discretized data

Data discretization is basically done to remove noisy data which may be inconsistent or incomplete so as to remove burden from machine. Data discretization changes the untidy value into discrete value. Here in default of credit card dataset the value of some of the attributes is very inconsistent, so to make them consistent data discretization is done on

attributes like Bill Amt1, Bill Amt2, Bill Amt3, Bill Amt4, Bill Amt5, Bill Amt6 as we can see that these attributes value consist of too much noisy data. For data discretization we divided the total instances into 10 equal size bin after arranging them in decreasing or increasing order. And then allotting numbering to those 10 bin like giving 1 to the first bin, 2 to second bin and so on up to 10^{th} bin. Figure 4.3 shows the data after discretization.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	LIMIT_BAL	SEX	EDUCATIC	MARRIAGI	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT:	BILL_AMT:	BILL_AMT:	BILL_AMT:	BILL_AMT:	BILL_AMT	PAY_AMT:
2	1	2	2	2	21	2	2	3	2	0	0	4	4	4	4	4	6	3905
3	1	1	3	2	36	0	0	0	0	0	0	4	4	4	4	4	6	1446
4	1	2	2	2	44	-2	-2	-2	-1	0	0	2	2	2	5	6	6	390
5	1	2	2	1	37	0	-1	0	0	0	0	4	4	4	4	4	6	7419
6	1	1	2	2	23	0	0	0	0	2	2	3	4	4	4	4	5	1117
7	1	2	2	2	42.40253	-2	-2	-2	-1	-0.39937	-0.39937	1	2	2	4	5	5	234.2464
8	1	2	3	1	33	2	2	2	0	0	0	4	4	4	4	4	5	2500
9	1	2	2	3	46	3	2	2	2	2	4	3	4	4	4	5	5	0
10	1	1	1	2	36	1	2	3	2	0	0	4	4	4	4	4	5	2620
11	1	2	2	1	23	1	3	2	2	2	0	4	4	4	5	5	5	0
12	1	2	2	1	23	1	3	2	2	2	0	4	4	4	5	5	5	0
13	1	1	3	2	36	-1	-1	-1	-1	2	2	4	2	1	4	4	5	344
14	1	2	1	1	39	1	-2	-1	2	0	0	1	1	4	4	4	5	0
15	1	2	2	2	31	0	0	0	0	0	0	4	4	4	4	5	5	1160
16	1	1	2	1	35	0	0	0	0	0	0	4	4	4	4	4	5	1160
17	1	1	2	1	35	0	0	0	0	0	0	4	4	4	4	4	5	1160
18	1	2	2	1	46	1	2	2	2	2	2	4	4	4	4	4	5	1306
19	1	1	2	2	21	0	0	0	0	0	0	4	4	4	4	5	5	1200
20	1	2	2	2	33	-2	-1	0	0	2	2	1	3	4	4	4	5	3000
21	1	1	2	2	27	2	2	2	0	0	2	4	4	4	4	4	5	2600
22	1	1	2	1	45	0	0	0	2	0	0	4	4	4	4	4	5	1400
23	1	2	2	2	32	0	0	0	0	0	0	4	4	4	4	4	5	1400

Figure 4.3: Discretized data

The table below shows the performance of various models on discretization data.

Table 4.3: Results of proposed method of Discretization data

Method	Accuracy	AUC	Error Rate	Sensitivity	Specificity
nnet	74.09	0.714	0.179	0.691	0.703
RF	78.84	0.743	0.201	0.761	0.850
Oblique	75.89	0.739	0.112	0.739	0.759
svm	76.10	0.659	0.144	0.749	0.781
Linear	75.17	0.691	0.253	0.831	0.859
Earth	77.11	0.729	0.179	0.788	0.897
Treebag	76.99	0.701	0.169	0.777	0.789

After comparing the results of both raw data and discretized data, it was obtained that there is no much difference in the result of both the data set. Discretization make the data clean and noiseless also discretization increases the performance. So it is considered better to perform discretization on default data set.

4.5 Features Selection

After performing various machine learning procedure on default of MasterCard dataset the assumption is that the performance on raw data is very close to the performance of discretized data. So it was thought to do discretization as it remove the noisy data as well as improve the performance.

Now if we talk about feature selection, here feature selection is done in a way that after removing various attributes from the dataset the performance remain the same on the basis of various parameters like sensitivity, specificity, accuracy, error rate, ROC etc. Also features selection makes the performance faster. Here feature selection is done using two technique namely Gini index and correlation.

4.5.1 Correlation Result

Firstly, the feature selection is done using correlation. After performing correlation on the normalize dataset of twenty four attributes, we noticed that three attributes namely BillAmt2, BillAmt3, BillAmt5 are least important so by removing these least important attribute we are left with twenty-one attributes. Figure 4.4 shows dataset after performing correlation on balanced data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	LIMIT_BAL	SEX	EDUCATIC	MARRIAG	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT	BILL_AMT	BILL_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT
2	1	2	2	2	21	2	2	3	2	0	0	4	4	4	3905	0	0	300	2165	2666
3	1	1	3	2	36	0	0	0	0	0	0	4	4	4	1446	2017	1000	186	190	0
4	1	2	2	2	44	-2	-2	-2	-1	0	0	2	5	6	390	780	17585	1000	667	692
5	1	2	2	1	37	0	-1	0	0	0	0	4	4	4	7419	1400	1000	500	500	500
6	1	1	2	2	23	0	0	0	0	2	2	3	4	4	1117	1700	1230	0	761	0
7	1	2	2	2	42.40253	-2	-2	-2	-1	-0.39937	-0.39937	1	4	5	234.2464	468.4928	10574.09	720.4423	799.9896	1214.374
8	1	2	3	1	33	2	2	2	0	0	0	4	4	4	2500	0	1200	500	1400	1
9	1	2	2	3	46	3	2	2	2	2	4	3	4	5	0	4000	0	2395	0	0
10	1	1	1	2	36	1	2	3	2	0	0	4	4	4	2620	0	0	400	1100	0
11	1	2	2	1	23	1	3	2	2	2	0	4	5	5	0	0	3000	0	500	1000
12	1	2	2	1	23	1	3	2	2	2	0	4	5	5	0	0	3000	0	500	1000
13	1	1	3	2	36	-1	-1	-1	-1	2	2	4	4	4	344	900	10040	0	790	0
14	1	2	1	1	39	1	-2	-1	2	0	0	1	4	4	0	9996	0	363	752	0
15	1	2	2	2	31	0	0	0	0	0	0	4	4	5	1160	1331	1321	517	528	600
16	1	1	2	1	35	0	0	0	0	0	0	4	4	4	1160	1139	1241	345	580	180
17	1	1	2	1	35	0	0	0	0	0	0	4	4	4	1160	1139	1241	345	580	180
18	1	2	2	1	46	1	2	2	2	2	2	4	4	4	1306	0	1735	0	2000	0
19	1	1	2	2	21	0	0	0	0	0	0	4	4	5	1200	1165	1079	1200	754	0
20	1	2	2	2	33	-2	-1	0	0	2	2	1	4	4	3000	3000	3000	0	2000	0
21	1	1	2	2	27	2	2	2	0	0	2	4	4	4	2600	0	1000	900	300	0
22	1	1	2	1	45	0	0	0	2	0	0	4	4	4	1400	1700	0	400	600	200
23	1	2	2	2	32	0	0	0	0	0	0	4	4	4	1400	1300	350	500	1650	0

Figure 4.4: Dataset after performing Correlation

The table below shows the performance of various models on correlation data.

Table 4.4: Results of proposed method on correlation data

Method	Accuracy	AUC	Error Rate	Sensitivity	Specificity
nnet	74.99	0.719	0.179	0.780	0.693
RF	77.14	0.741	0.199	0.771	0.832
Oblique	78.81	0.796	0.112	0.723	0.779
svm	76.99	0.655	0.142	0.745	0.761
Linear	75.88	0.645	0.143	0.801	0.869
Earth	77.81	0.769	0.199	0.820	0.877
Treebag	74.77	0.721	0.169	0.677	0.759

From the correlation performance table we can see that there is not much difference in the performance of normalized data and correlation. And also removing the extra attributes makes the performance faster.

4.5.2 Gini Index Result

Secondly the Gini index technique is performed on the correlation data i.e the data left with twenty-one attributes. It was noted that the six attribute are least importance in the dataset namely SEX, EDUCATION, MARRIAGE, PAY_6, BillAmt4, BillAmt6. So after removing these six attributes we are left with fifteen important attributes. The following figure shows the dataset after performing Gini index technique.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	LIMIT_BAL	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	BILL_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	default.payment.next.month		
2	1	21	2	2	3	2	0	4	3905	0	0	300	2165	2666	1		
3	1	36	0	0	0	0	0	4	1446	2017	1000	186	190	0	0		
4	1	44	-2	-2	-2	-1	0	2	390	780	17585	1000	667	692	1		
5	1	37	0	-1	0	0	0	4	7419	1400	1000	500	500	500	0		
6	1	23	0	0	0	0	2	3	1117	1700	1230	0	761	0	0		
7	1	42.40253	-2	-2	-2	-1	-0.39937	1	234.2464	468.4928	10574.09	720.4423	799.9896	1214.374	1		
8	1	33	2	2	2	0	0	4	2500	0	1200	500	1400	1	0		
9	1	46	3	2	2	2	2	3	0	4000	0	2395	0	0	1		
10	1	36	1	2	3	2	0	4	2620	0	0	400	1100	0	1		
11	1	23	1	3	2	2	2	4	0	0	3000	0	500	1000	0		
12	1	23	1	3	2	2	2	4	0	0	3000	0	500	1000	0		
13	1	36	-1	-1	-1	-1	2	4	344	900	10040	0	790	0	0		
14	1	39	1	-2	-1	2	0	1	0	9996	0	363	752	0	1		
15	1	31	0	0	0	0	0	4	1160	1331	1321	517	528	600	0		
16	1	35	0	0	0	0	0	4	1160	1139	1241	345	580	180	0		
17	1	35	0	0	0	0	0	4	1160	1139	1241	345	580	180	0		
18	1	46	1	2	2	2	2	4	1306	0	1735	0	2000	0	1		
19	1	21	0	0	0	0	0	4	1200	1165	1079	1200	754	0	0		
20	1	33	-2	-1	0	0	2	1	3000	3000	3000	0	2000	0	0		
21	1	27	2	2	2	0	0	4	2600	0	1000	900	300	0	1		
22	1	45	0	0	0	2	0	4	1400	1700	0	400	600	200	1		
23	1	32	0	0	0	0	0	4	1400	1300	350	500	1650	0	0		

Figure 4.5: Dataset after performing Gini index

The table below shows the performance of various model on Gini index dataset.

Table 4.5: Results of proposed method on Gini index data

Method	Accuracy	AUC	Error Rate	Sensitivity	Specificity
nnet	74.09	0.721	0.186	0.771	0.733
RF	78.14	0.753	0.219	0.771	0.862
Oblique	77.89	0.736	0.122	0.733	0.769
svm	76.72	0.64	0.14	0.755	0.771
Linear	75.18	0.631	0.243	0.801	0.869
Earth	77.01	0.739	0.177	0.828	0.877
Treebag	76.77	0.711	0.172	0.777	0.789

After comparing the performance of both the correlation dataset and the Gini index dataset, it was noticed that there is no much difference in the result of both the techniques. Also the result of Gini index is quite better than that of correlation dataset. So we decided to continue with the Gini index dataset.

4.6 Comparison on Different Partition

As we have already discussed that the dataset is parted into two partition i.e. namely in testing and training. So till now the results were obtained on the partition of 80-20 i.e. 80% training and 20% testing. Now for the purpose of comparing the data we will partition the data with different ratio and taking the result of all the different partition and then compare the result. There are two parameters for computing the performance namely accuracy and ROC value. This work is basically done to check that whether the models are functioning properly or not for every partition.

4.6.1 Accuracy on Different Partition

In table 4.6 the data is partitioned into four partitions i.e. 90-10, 80-20, 70-30, 60-40. Now the evaluation is done on various partitions. Firstly for Random forest when partition is 90-10 accuracy is 82.10, when partition is 80-20 accuracy is 77.69, which is very near to prior partition, when partition is 70-30 accuracy is 78.21, which is also very near to both the prior result, similarly when partition is 60-40 accuracy is 76.95, which is also very near to all prior result. Similarly for Stochastic gradient boosting model when partition is 90-10 accuracy is 80.75, when partition is 80-20 accuracy is 79.76, which is very near to prior result, when partition is 70-30 accuracy is 78.82, which is very close to both prior result, when partition is 60-40 accuracy is 77.71, which is very much near to all the prior result. Similarly for multivariate adaptive spline when partition is 90-10 accuracy is 78.52, when partition is 80-20 accuracy is 77.81, when partition is 70-30 accuracy is 76.97, which is very near to both the prior result, when partition is 60-40 accuracy is 76.03, which is very near to all the prior result. Similarly, for conditional inference tree model when dataset is partitioned into 90-10, 80-20, 70-30, 60-40 accuracies are 78.70, 77.29, 76.29, 75.08 respectively where all the accuracies are too much near to each other. Similarly, for decision tree model, earth model and treebag the correctness for various partitions are near to each other that means prototypes are performing good on various partitions. This shows that many more machine models can be executed and best model can be predicted on this dataset. The table below shows accuracy on different partition.

Table 4.6 : Accuracy on Different Partition

S.no	Model Name	Partition1	Partition2	Partition3	Partition4
		90-10	**80-20**	**70-30**	**60-40**
1	**RF**	82.10	77.69	78.21	76.95
2	**Gbm**	80.75	79.76	78.82	77.71
3	**Earth**	78.52	77.81	76.97	76.03
4	**Ctree**	78.70	77.29	76.29	75.08
5	**Rpart**	77.61	76.77	76.10	75.63

4.6.2 ROC on Different Partition

In table 4.7 the data is partitioned into four partitions i.e. 90-10, 80-20, 70-30, 60-40. Now the evaluation is completed on various partitions. Firstly for Random forest when partition is 90-10 ROC is 82.10, when partition is 80-20 ROC is 77.69, which is very near to prior partition, when partition is 70-30 ROC is 78.21, which is also very near to both the prior result, similarly when partition is 60-40 ROC is 76.95, which is also very near to all prior result. Similarly for Stochastic gradient boosting model when partition is 90-10 ROC is 80.75, when partition is 80-20 ROC is 79.76, which is very near to prior result, when partition is 70-30 ROC is 78.82, which is very close to both prior result, when partition is 60-40 ROC is 77.71, which is very much near to all the prior result. Similarly for multivariate adaptive spline when partition is 90-10 ROC is 78.52, when partition is 80-20 ROC is 77.81, when partition is 70-30 ROC is 76.97, which is very near to both the prior result, when partition is 60-40 ROC is 76.03, which is very near to all the prior result. Similarly, for conditional inference tree model when dataset is partitioned into 90-10, 80-20, 70-30, 60-40 ROC values are 78.70, 77.29, 76.29, 75.08 respectively where all the ROC are too much near to each other. Similarly, for decision tree model, earth model and treebag the ROC for various partitions are near to each other that means prototypes are performing good on various partitions. This shows that many more machine models can be executed and best model can be predicted on this dataset. The table below shows ROC on different partition.

Table 4.7: ROC on various Partition

S.no	Model Name	Partition1	Partition2	Partition3	Partition4
		90-10	**80-20**	**70-30**	**60-40**
1	**RF**	0.801	0.7391	0.764	0.745
2	**Gbm**	0.811	0.781	0.783	0.783
3	**Earth**	0.730	0.729	0.724	0.713
4	**Ctree**	0.746	0.749	0.722	0.713
5	**Rpart**	0.595	0.697	0.618	0.626

4.7 Model Comparison

After relating the presentation of various partitions, next job is to relate the dataset for various models. Models utilized are Mixture Discriminant Analysis, Neural net, Random forest, Boosted Classification tree, Stochastic Gradient Boosting, Oblique tree, Decision Tree, Support vector machine, Multivariate Adaptive Spline, conditional inference tree, Treebag, Generalized Linear Modal, Linear model, Flexible discriminant analysis, Partial Least Square. On the basis of various parameters like Sensitivity, Specificity, Error rate, ROC and Accuracy comparison is done. The comparison is performed on fifteen features as shown in table 4.8. Now top five prototype of the dataset out of fifteen prototype will be selected. After examining the value of all the parameters like ROC, Accuracy, Sensitivity, Specificity and Error Rate for fifteen dissimilar prototype. Top five prototypes for the default dataset are neural network, Random forest, Oblique Tree, SVM, linear, Earth, Treebag.

Table 4.8: Comparison of Fifteen Models

Method	Accuracy	AUC	Error Rate	Sensitivity	Specificity
RF	82.11	0.819	0.199	0.782	0.856
Gbm	79.76	0.781	0.202	0.794	0.801
Earth	77.81	0.729	0.156	0.824	0.890
Ctree	77.29	0.749	0.127	0.774	0.821
Rpart	76.77	0.697	0.232	0.775	0.791
Treebag	76.75	0.718	0.182	0.775	0.795
Oblique	76.66	0.726	0.099	0.733	0.789
Svm	76.42	0.690	0.137	0.765	0.771
Fda	75.97	0.801	0.147	0.783	0.817
Linear	75.17	0.687	0.212	0.811	0.846
Nnet	74.97	0.720	0.174	0.762	0.724
Glm	74.87	0.732	0.109	0.779	0.851
Pls	72.13	0.717	0.223	0.754	0.857
Ada	74.17	0.701	0.089	0.766	0.737
Mda	71.47	0.689	0.169	0.759	0.799

4.8 K-Fold

After sorting the best five models out of fifteen models for the dataset, our next job is to perform k-fold on that dataset. K-fold validation is utilized for measuring the power of predictive model. In this approach information is divided into k square with size out of the k test single example is split with as validation ensemble model information. At that point cross validation approach is done on k times with every k sub test utilized precisely once as a validation information [29]. Cross Validation is done on best five models namely random forest, Gbm, Earth, Ctree, Rpart where the value of k is taken as 10.

4.8.1 K-fold on Random Forest

Carrying out k-fold validation on Random Forest prototype. Performing 10-fold means to run the model ten times on shuffled data and then noting down the value of accuracy. Table 4.9 shows the value of accuracy obtained after performing the k-fold validation ten times. It can be clearly seen in table that the accuracy obtained are very much closed to one another.

Table 4.9: K-Fold on Random Forest

S.No.	Accuracy
1	82.117
2	83.099
3	82.991
4	81.137
5	83.773

S.No.	Accuracy
6	82.527
7	83.093
8	80.994
9	82.021
10	83.010

The plot for Random Forest prototype is shown in Figure 4.6 for ten repetitions. The plot of the graph shows the robustness of the prototype.

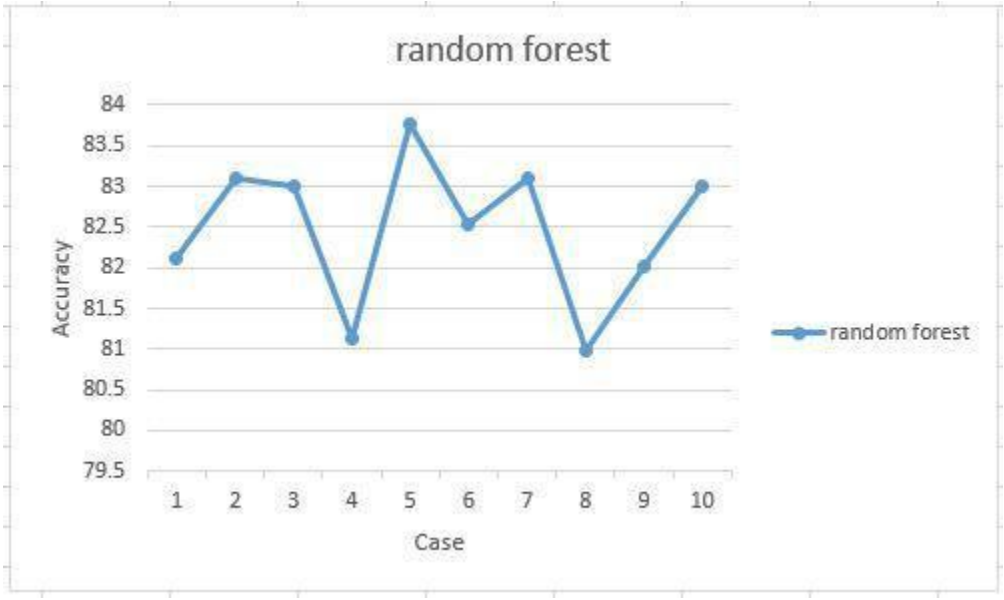


Figure 4.6: Performance for Random Forest Model

4.8.2 K-Fold on Stochastic Gradient Boosting

Carrying out k-fold validation on Stochastic Gradient Boosting prototype. Performing 10-fold means to run the model ten times on shuffled data and then noting down the value of accuracy. Table 4.10 shows the value of accuracy obtained after performing the k-fold validation ten times. It can be clearly seen in table that the accuracy obtained are very much closed to one another.

Table 4.10: K-Fold on Stochastic Gradient Boosting

S.No.	Accuracy
1	79.760
2	79.015
3	80.115
4	78.233
5	80.456

S.No.	Accuracy
6	79.137
7	78.542
8	80.029
9	79.527
10	80.514

The plot for Stochastic Gradient Boosting prototype is shown in Figure 4.7 for ten repetitions. The plot of the graph shows the robustness of the prototype.

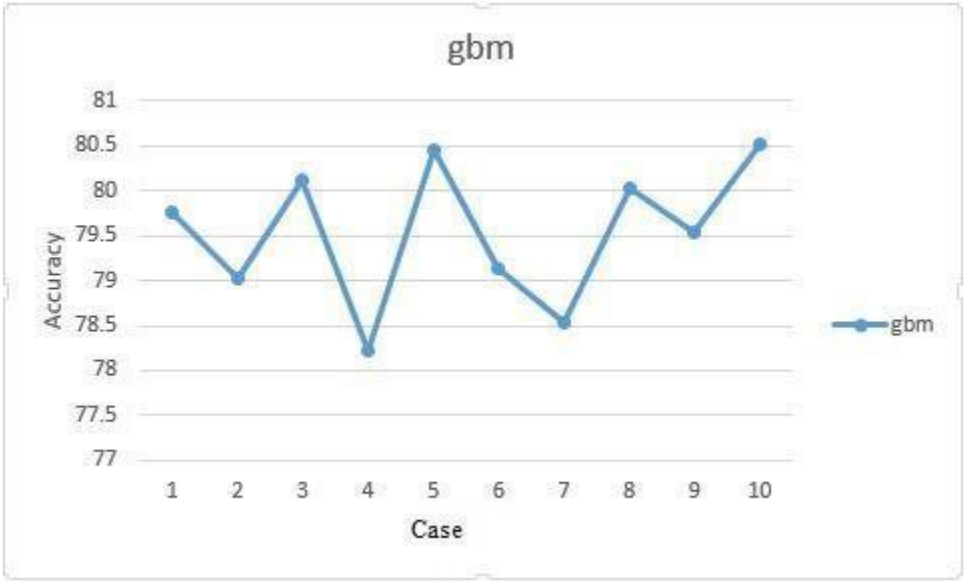


Figure 4.7: Performance for Stochastic Gradient Boosting.

4.8.3 K-Fold on Multivariate Adaptive Spline

Carrying out k-fold validation on Multivariate Adaptive Spline prototype. Performing 10-fold means to run the model ten times on shuffled data and then noting down the value of accuracy. Table 4.11 shows the value of accuracy obtained after performing the k-fold validation ten times. It can be clearly seen in table that the accuracy obtained are very much closed to one another.

Table 4.11: K-Fold on Multivariate Adaptive Spline

S.No.	Accuracy
1	77.813
2	78.990
3	77.236
4	79.458
5	76.392

S.No.	Accuracy
6	77.921
7	79.767
8	78.239
9	77.527
10	78.010

The plot for Multivariate Adaptive Spline prototype is shown in Figure 4.8 for ten repetitions. The plot of the graph shows the robustness of the prototype.

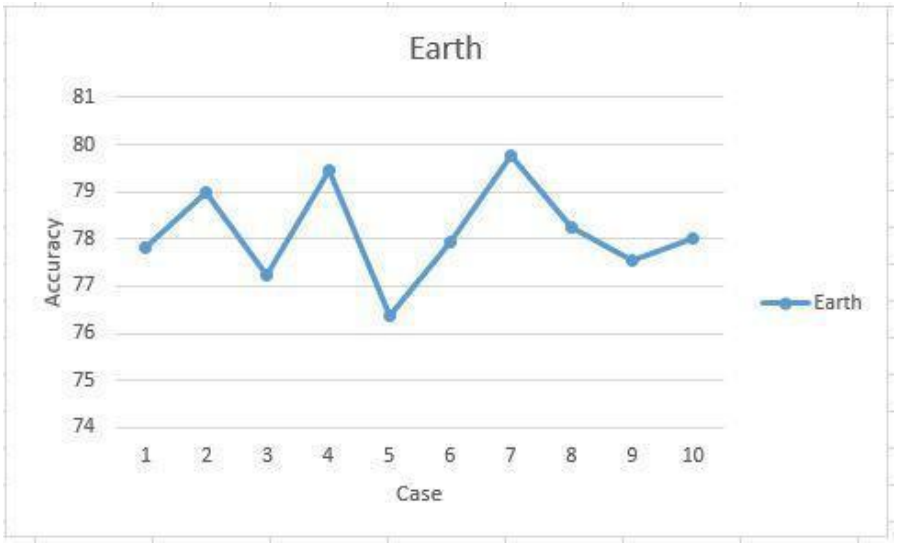


Figure 4.8: Performance for Multivariate Spline Model

4.8.4 K-Fold on Conditional Inference Tree

Carrying out k-fold validation on Conditional Inference Tree prototype. Performing 10-fold means to run the model ten times on shuffled data and then noting down the value of accuracy. Table 4.12 shows the value of accuracy obtained after performing the k-fold validation ten times. It can be clearly seen in table that the accuracy obtained are very much closed to one another.

Table 4.12: K-Fold on Conditional Inference Tree

S.No.	Accuracy
1	77.296
2	76.992
3	76.997
4	78.471
5	76.973

S.No.	Accuracy
6	77.391
7	78.679
8	77.021
9	76.238
10	77.781

The plot for Conditional Inference Tree prototype is shown in Figure 4.9 for ten repetitions. The plot of the graph shows the robustness of the prototype.

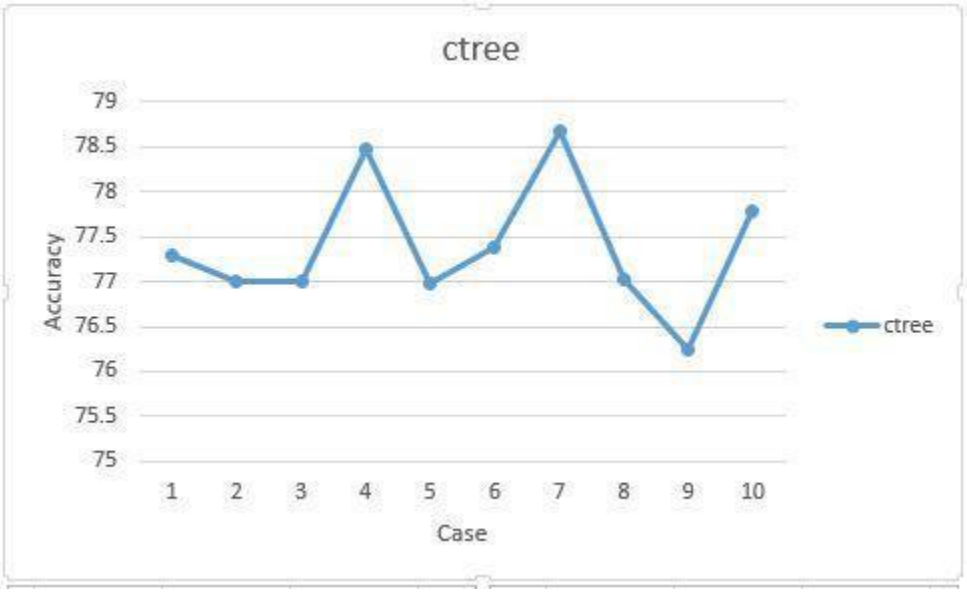


Figure 4.9: Performance for Conditional Inference Tree

4.8.5 K-Fold on Decision Tree

Carrying out k-fold validation on Decision Tree prototype. Performing 10-fold means to run the model ten times on shuffled data and then noting down the value of accuracy. Table 4.13 shows the value of accuracy obtained after performing the k-fold validation ten times. It can be clearly seen in table that the accuracy obtained are very much closed to one another.

Table 4.13: K-Fold on Decision Tree

S. No.	Accuracy
1	76.771
2	80.492
3	75.993
4	77.234
5	76.595

S. No.	Accuracy
6	77.298
7	76.456
8	78.011
9	77.489
10	76.321

The plot for Decision Tree prototype is shown in Figure 10 for ten repetitions. The plot of the graph shows the robustness of the prototype.

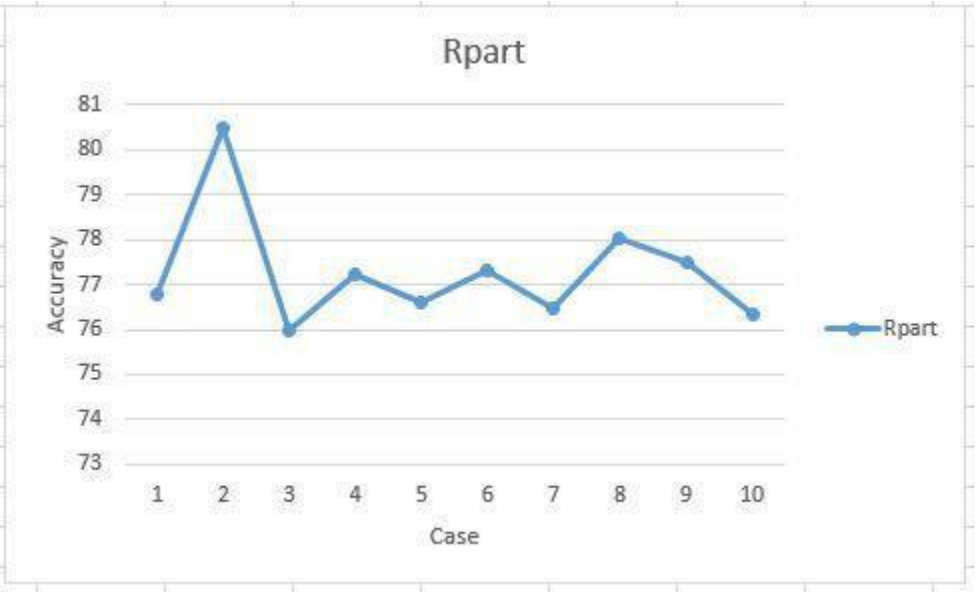


Figure 4.10: Performance for Decision Tree

4.9 Ensemble

After comparing all the machine learning model performed, our next job is to select best five models for default dataset. After carrying out comparison based on various machine learning parameters such as Accuracy, Error Rate, Specificity, Sensitivity and ROC value (Table 4.8) top five models are random forest, stochastic gradient boosting, multivariate adaptive spline, conditional inference tree, decision tree for default MasterCard customer dataset.

Now next job is to group the best five models in such a manner that gives out maximum correctness. Table 4.14 shows the testing executed on best five models and for the best five models we performed various permutation to compute highest correctness. After executing twenty permutation, highest accuracy is 80.533 which is for random forest and stochastic gradient boosting. Now on comparing the accuracy of the ensemble model (table 4.14) and the accuracy obtained by individual model (table 4.8), it is noticed that ensemble model accuracy is much better than the individual model accuracy.

After executing ensemble on every permutation on the best five model i.e. random forest, stochastic gradient boosting, multivariate adaptive spline, conditional inference tree, decision tree. We can see easily see from the table 4.14 that the top ensemble prototype for the default of MasterCard customer dataset is random forest, stochastic gradient boosting with accuracy 80.533. Figure 4.6 to 4.10 shows individual chart for best five models for default of MasterCard dataset. Figure 4.11 shows a collective chart that is achieved after accomplishment of model ensemble.

Table 4.14: Ensemble Models

S. No.	Ensemble Model	Accuracy
1	rf, gbm	80.533
2	rf,earth	78.77
3	rf,ctree	78.31
4	rf,rpart	77.86
5	rf,gbm,earth	78.45
6	rf,gbm,ctree	78.05
7	rf,gbm,rpart	77.70
8	rf,gbm,earth,ctree	77.67
9	rf,gbm,earth,rpart	78.64
10	rf,gbm,earth,ctree,rpart	77.67
11	gbm,earth	79.37
12	gbm,ctree	79.29
13	gbm,rpart	79.16
14	gbm,earth,ctree	78.57
15	gbm,earth,rpart	78.31
16	gbm,earth,ctree,rpart	78.24
17	earth,ctree	78.01
18	earth,rpart	77.50
19	earth,ctree,rpart	77.06
20	ctree,rpart	77.48

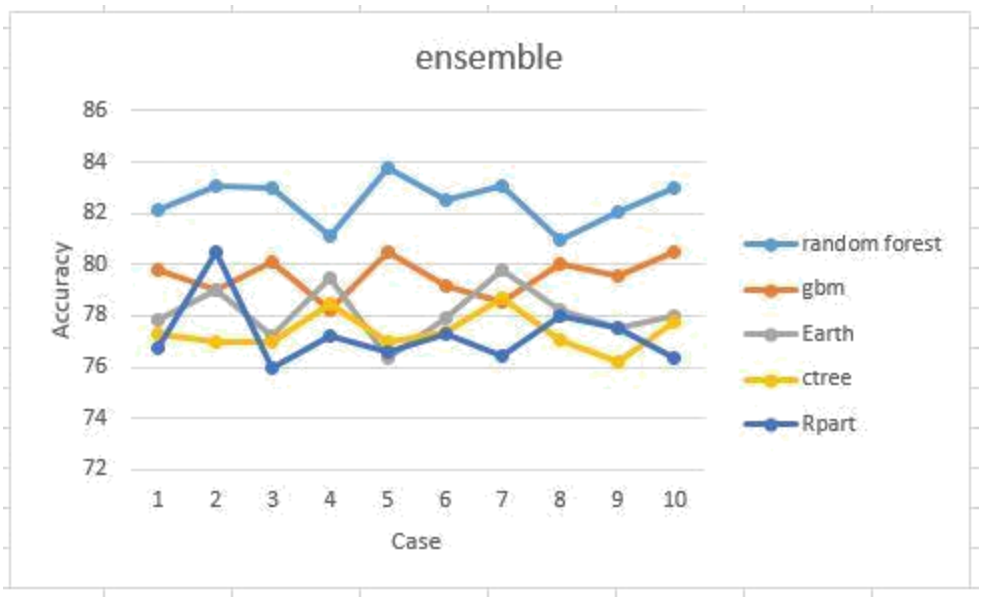


Figure 4.11: Ensemble Model Plot

Chapter-5

Conclusion and Future Scope

This section discusses the involvement of thesis and coming work of this proposal and also defines conclusion of the effort present in this work.

5.1 Conclusion

The significances of our procedure have established that ensemble model of Random Forest and Stochastic Gradient Boosting are as one the top combination to develop the desired consequence effectively.

In this procedure we performed our task on fifteen different models for default of MasterCard customer dataset. After executing fifteen models on the default of MasterCard dataset, the best five models for the default dataset on the basis of various parameters like ROC value, accuracy, specificity, sensitivity and error rate are random forest, stochastic gradient boosting, multivariate adaptive spline, conditional inference tree and decision tree. Various permutation were performed on the best five models (table 4.14). After executing all the permutation, group model for default of MasterCard customer dataset is recommended.

5.2 Summary of Contributions

The influence done for the study presented in this theory are shortened as follows:

a) Firstly, the dataset was collected from UCI repository.
b) Model were compared on the basis of accuracy, ROC curve, error rate, specificity and sensitivity.
c) Procedure to transform output predicted from the model into accuracy using excel sheet.
d) Relative examination of various models based on different partition.

5.3 Future Scope

To improve the accuracy or correctness of the prototypes ensembling plays a great role. This task is been verified on default of MasterCard dataset. Different process utilized for default dataset are random forest, rreebag, ctree, linear model etc. Instead of using the individual algorithm, if default of MasterCard dataset practices planned ensemble model approach than after ensembling default of MasterCard becomes more effective.

References

[1] I-Cheng Yeh and Che-hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", Expert Systems with Applications, 36(2):2473–2480, 2009.

[2] Little and Ken, "Personal Finance At Your Fingertips" , p. 35 Penguin 2007. ISBN 144062562X, 9781440625626

[3] "Report to the Congress on the Use of Credit Cards by Small Businesses and the Credit Card Market for Small Businesses"(PDF). Federal Reserve. Board of Governors of the Federal Reserve System. May 2010. Retrieved 4 May 2015.

[4] "Credit Cards and You – About Pre-paid Cards". Financial Consumer Agency of Canada. Archived from the original on 7 March 2007. Retrieved 9 January 2008. document: "Pre-paid Cards" (PDF). Financial Consumer Agency of Canada. Archived from the original (PDF) on 29 February 2008. Retrieved 9 January 2008.

[5] Federal Reserve Bank of Kansas City, The Economics of Payment Card Fee Structure: What Drives Payment Card Rewards?, March 2009

[6] CreditCards.com (27 January 2010). "Credit card penalty rates can top 30 percent; how to avoid them". Creditcards.com. Retrieved 26 March 2013.

[7] Drazen Prelec & George Loewenstein (21 December 1998), "The Red and the Black: Mental Accounting of Savings and Debt", Mktsci.journal.informs.org. Archived from the original on 10 July 2012. Retrieved 26 March 2013.

[8] "Finally, Money Advice That Will Make You Skinnier". Time. 7 July 2011.

[9] Martin and Andrew (4 January 2010), "How Visa, Using Card Fees, Dominates a Market", New York Times. Retrieved 6 January 2010. The fees, roughly 1 to 3 percent of each purchase, are forwarded to the cardholder’s bank to cover costs and promote the issuance of more Visa cards.

[10] Dickler and Jessica (31 July 2008), "Hidden credit card fees are costing you", CNN. Retrieved 30 April 2010.

[11] Gensler and Lauren (April 2013), "You (Probably) Won't Pay More to Swipe", Money, New York, p. 14

[12] Douglas and Danielle, "Judge approves Visa, MasterCard $5.7 billion settlement with retailers", Washington Post.

[13] "PCI Compliance", Thrive Business Solutions. Archived from the original on 5 March 2008.

[14] Credit Card Issuer Fraud Management, Report Highlights, December, 2008

[15] The Interchange Debate: Issues and Economics James Lyon, 19 January 2006 Archived 22 March 2008 at the Wayback Machine.

[16] M. Lichman, UCI machine learning repository, 2013.

[17] SB Kotsiantis, D Kanellopoulos, and PE Pintelas, Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2):111–117, 2006.

[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "Smote: synthetic minority oversampling technique", Journal of artificial intelligence research, 16:321–357, 2002.

[19] Thomas Grubinger, Achim Zeileis, and Karl-Peter Pfeiffer, "evtree: Evolutionary learning of globally optimal classification and regression trees" in r. Technical report, Working Papers in Economics and Statistics, 2011.

[20] Jerome H Friedman, "Multivariate adaptive regression splines", The annals of statistics, pages 1–67, 1991.

[21] Yoav Freund and Robert E Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting", in European conference on computational learning theory, pages 23–37. Springer, 1995.

[22] Trevor Hastie, Robert Tibshirani, and Andreas Buja, "Flexible discriminant analysis by optimal scoring", Journal of the American statistical association, 89(428):1255–1270, 1994.

[23] Martin Riedmiller and Heinrich Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm", in Neural Networks, 1993., IEEE International Conference on, pages 586–591. IEEE, 1993.

[24] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers", Foundations and Trends R in Machine Learning, 3(1):1–122, 2011.

[25] J. Ross Quinlan, "Induction of decision trees", Machine learning, 1(1):81–106, 1986.

[26] Andy Liaw and Matthew Wiener, "Classification and regression by randomforest", R news, 2(3):18–22, 2002.

[27] S. Sathiya Keerthi and Elmer G Gilbert, "Convergence of a generalized smo algorithm for svm classifier design", Machine Learning, 46(1-3):351–360, 2002.

[28] Richard A Becker and John M Chambers," S: an interactive environment for data analysis and graphics", CRC Press, 1984.

[29] Payam Refaeilzadeh, Lei Tang, and Huan Liu, "Crossvalidation", in Encyclopedia of database systems, pages 532–538. Springer, 2009.

List of Publications

[1] Vaishali and Rajkumar Tekchandani, "Default of MasterCard Customers Prediction using Machine Learning Approaches", International Conference on Intelligent Computing and Control (I2C2) 2017.
[Accepted]

Video URL

https://youtu.be/7AwDlt892s4