IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 11, PP 05-09

Credit Risk Analysis

Harshala Yadav[1], Hemani Yadav[2], Divya Kumawat[3]
1, 2 (Student ,Department Of Computer Engineering, Atharva College Of Engineering, Mumbai University , India)
3 (Assistant Professor, Department Of Computer Engineering, Atharva College Of Engineering, Mumbai University , India)

Abstract: ***Nowadays*** *There Is Huge Data That Is Being Used By Organizations And It Is Required To Extract Important Patterns And Information From The Database That Is Useful. For This Purpose Data Mining Techniques Is Used. Decision Tree Is Very Popular For Classification Purpose In Data Mining. As We Know In Banking Sector Huge Databases Are Used For Credit Risk Assessment .Credit Risk Is Any Loss That Can Be Incurred Due To Debt Defaulters Encountered. For This Purpose Data Mining Techniques Like Decision Tree Is Very Efficient For Prediction Of Defaulters Which Will Lower The Risk Associated With Lending Loans.*
Keywords *-Banking Industry, Credit Scoring, Data Mining, Decision Tree*

I. Introduction

Nowadays The Traditional Way For Customer Contacts Is Replaced By Electronic Way So As To Reduce The Time Taken To Verify The Customers Details. The Credit Function Is Important Part Of Banking. Interest Income Is The Main Source Of Income For Any Bank. Risk Is Part Of Bank's Business. Lending Any Loan To Customer Always Involves Some Risk. Data Mining Can Be Used To Find Patterns And Connections That Would Otherwise Be Difficult To Find. This Technology Is Popular With Many Businesses Because It Allows Them To Learn More About Their Customers And Make Smart Marketing Decisions [1].

There Are Many Risks Related To Bank Loans But The Major Risks Lie In Lending Loans .The Most Important Thing Is To Analyze The Defaulters Or Customers Who Can Be Unable To Pay The Loan. The Analysis Of Risks Becomes Crucial Thereafter. Banks Have Large Customer Data From Which They Are Unable To Arrive At A Conclusion. Data Mining Is A Promising Area Of Data Analysis Which Extracts Useful Knowledge From Tremendous Amount Of Complex Data Sets. .

The Existing Model Is Built Using The Data Mining Functions Available In The R Package And Dataset Is Taken From The UCI Repository. As The Pre- Processing Step Is The Most Important And Time Consuming One, Classification And Clustering Techniques In R Were Used To Make The Data Ready For Further Use. Pre-Processed Data Set Is Then Used For Building The Decision Tree Classifier. The Tree Model Is Then Used To Predict The Class Labels Of The New Loan Applicants, Their Probability Of Default. Several R Functions And Packages Were Used To Prepare Data And To Build The Classification Model [2].

The Main Objective Of Our Model Is To Classify The Loan Based On Credit Scoring And Behavioral Analysis. Credit Scoring Decides Whether To Grant Loan To The Customer. Assessing A Customers Behavior Is Also Important So That We Can Lend Loan To The Right Applicant. Data Mining Technique Is Used For This Purpose. It Is Used For Extraction Of Information From Large Database. Decision Tree Induction Algorithm Is One Of The Techniques To Achieve This Objective. The Model Thus Developed Will Provide A Credit Risk Assessment, Which Will Lead To Better Allocation Of The Capital.

What Is 'Credit Risk'?

Credit Risk Refers To The Risk That A Borrower May Not Repay A Loan And That The Lender May Lose The Principal Of The Loan Or The Interest Associated With It. Credit Risk Arises Because Borrowers Expect To Use Future Cash Flows To Pay Current Debt [3].

II. Literature Review

Mrs. Bharati M. Ramageri Discusses Few Of The Data Mining Techniques, Algorithms And Some Of The Organizations Which Have Adapted Data Mining Technology To Improve Their Businesses And Found Excellent Results [1].

Dr. K. Chitra , B. Subashini Analyzes The Data Mining Techniques And Its Applications In Banking Sector Like Fraud

Prevention And Detection, Customer Retention, Marketing And Risk Management. They Use Data Warehousing To

Combine Various Data From Databases Into An Acceptable Format So That The Data Can Be Mined. The Data Is Then

Analyzed And The Information That Is Captured Is Used Throughout The Organization To Support Decision-Making.[5]

Sudhamathy G., Jothi Venkateswaran C. Presented A Framework To Effectively Identify The Probability Of Default

Of A Bank Loan Applicant. Probability Of Default Estimation Can Help Banks To Avoid Huge Losses. This Model Is Built Using The Data Mining Functions Available In The R Package And Data Set Is Taken From The UCI Repository.[2]

Qasem A. Al-Radaideh ,Eman Al Nagi Conducted A Study Where Data Mining Techniques Were Utilized To Build A Classification Model To Predict The Performance Of Employees. To Build The Classification Model The CRISP-DM Data Mining Methodology Was Adopted. Decision Tree Was The Main Data Mining Tool Used To Build The Classification Model, Where Several Classification Rules Were Generated.[8]

2.1. Data Mining

The Development Of Information Technology Has Generated Large Amount Of Databases And Huge Data In Various Areas. The Research In Databases And Information Technology Has Given Rise To An Approach To Store And Manipulate This Precious Data For Further Decision Making. Data Mining Is A Process Of Extraction Of Useful Information And Patterns From Huge Data. It Is Also Called As Knowledge Discovery Process, Knowledge Mining From Data, Knowledge Extraction Or Data /Pattern Analysis.[1]

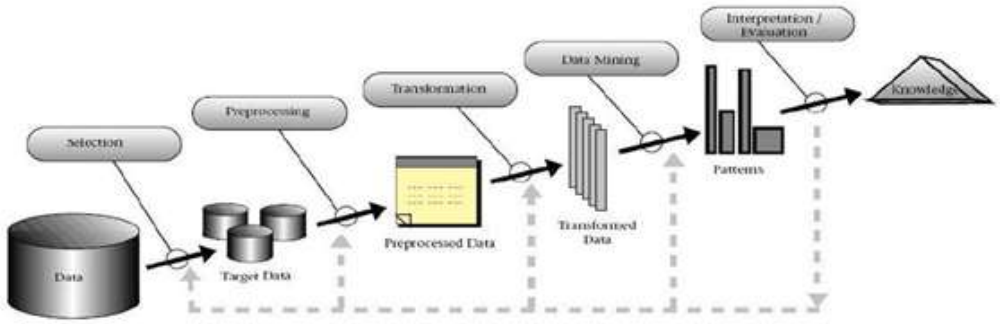

Fig.1 Knowledge Discovery Process

Data Mining Is A Logical Process That Is Used To Search Through Large Amount Of Data In Order To Find Useful Data. The Goal Of This Technique Is To Find Patterns That Were Previously Unknown. Once These Patterns Are Found They Can Further Be Used To Make Certain Decisions For Development Of Their Businesses.[1]

Data Mining Algorithms And Techniques Various Algorithms And Techniques Like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor Method Etc., Are Used For Knowledge Discovery From Databases[1].

2.2 Classification

Classification Consists Of Predicting A Certain Outcome Based On A Given Input. In Order To Predict The Outcome, The Algorithm Processes A Training Set Containing A Set Of Attributes And The Respective Outcome, Usually Called Goal Or Prediction Attribute. The Algorithm Tries To Discover Relationships Between The Attributes That Would Make It Possible To Predict The Outcome [4].

Classification Is The Most Commonly Applied Data Mining Technique, Which Employs A Set Of Pre - Classified Examples To Develop A Model That Can Classify The Population Of Records At Large. Fraud Detection And Credit Risk Applications Are Particularly Well Suited To This Type Of Analysis. This Approach Frequently Employs Decision Tree Or Neural Network-Based Classification Algorithms. The Data Classification Process Involves Learning And Classification [1].

2.3 Decision Tree

Decision Trees Are The Most Popular Predictive Models (Burez And Van Den Poel, 2007). A Decision Tree Is A Tree-Like Graph Representing The Relationships Between A Set Of Variables. Decision Tree Models Are Used To Solve Classification And Prediction Problems Where Instances Are Classified Into One Of Two Classes, Typically Positive

And Negative, Or Churner And Non-Churner In The Churn Classification Case. These Models Are Represented And Evaluated In A Top-Down Manner [5].

A Decision Tree Is A Flowchart-Like Structure In Which Each Internal Node Represents A "Test" On An Attribute .Each Branch Represents The Outcome Of The Test, And Each Leaf Node Represents A Class Label .The Paths From Root To Leaf Represent Classification Rules[6].

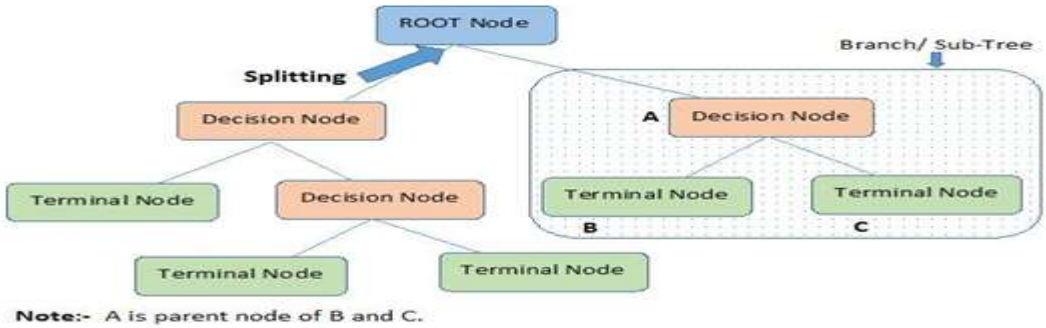

Fig.2 Decision Tree

Basic Terminology Used With Decision Trees:

- Root Node**:** It Represents Entire Population Or Sample And This Further Gets Divided Into Two Or More Homogeneous Sets.
- Splitting**:** It Is A Process Of Dividing A Node Into Two Or More Sub-Nodes.
- Decision Node: When A Sub-Node Splits Into Further Sub-Nodes, Then It Is Called Decision Node.
- Leaf/ Terminal Node: Nodes Do Not Split Is Called Leaf Or Terminal Node.
- Pruning: When We Remove Sub-Nodes Of A Decision Node, This Process Is Called Pruning. You Can Say Opposite Process Of Splitting.
- Branch / Sub-Tree: A Sub Section Of Entire Tree Is Called Branch Or Sub-Tree.
- Parent And Child Node: A Node, Which Is Divided Into Sub-Nodes Is Called Parent Node Of Sub-Nodes Where As Sub-Nodes Are The Child Of Parent Node [7].

III. Research Methodology

The CRISP- DM Methodology (Cross Industry Standard Process For Data Mining) (CRISP- DM, 2007) Was Used To Build A Classification Model. It Consists Of Five Steps Which Include: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation [8].

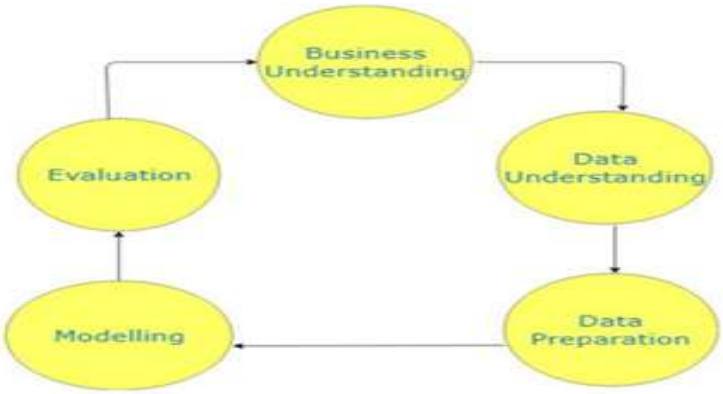

Fig.3 Crisp- Model

Business Understanding: Focuses On Understanding The Project Objectives And Requirements From A Business Perspective, Then Converting This Knowledge Into A Data Mining Problem Definition.

Data Understanding: Starts With An Initial Data Collection And Proceeds With Activities In Order To Get Familiar With The Data, To Identify Data Quality Problems.
Data Preparation: Covers All Activities To Construct The Final Data Set From The Initial Raw Data. Data Preparation Tasks Are Likely To Be Performed Multiple Times And Not In Any Prescribed Order.
Modeling: Various Modeling Techniques Are Selected And Applied And Their Parameters Are Calibrated To Optimal Values.
Evaluation: Determine If Results Meet Business Objectives; Identify Business Issues That Should Have Been Addressed Earlier [9].

IV. Proposed Model

The Credit Risk Assessment Is Very Crucial While Approving A Loan To Customer. Credit Risks Accounts For The Risk And Loan Defaults Which Is The Major Source Of Risk Encountered By Banking Industry. In Existing System The Objective Is To Develop A Data Mining Model Using R For Predicting PD For New Loan Applicants Of A Bank. The Data Used To Implement And Test This Model Is Taken From The UCI Repository. The German Credit Scoring Data Set With 1000 Records And 21 Attributes Is Used For This Purpose [2].

To Prevent Defaulters We Have Proposed A Two Step Credit Risk Assessment Model Which Will Identify Which Customer Will Be Able To Pay The Loan. The Proposed Model Focuses On Classifying Customer Loan Requests By Analyzing Their Data. The Model Takes Customer Information As Input, The Output Is Given By Decision Tree Which Predicts The Credibility Of Customer. The Proposed System Will Take Inputs Without Any Null Values.

The Proposed Model Focuses On Classifying Customer Loan Requests By Analyzing Their Data. The Model Takes Customer Information As Input, The Output Is Given By Decision Tree Which Predicts The Credibility Of Customer.

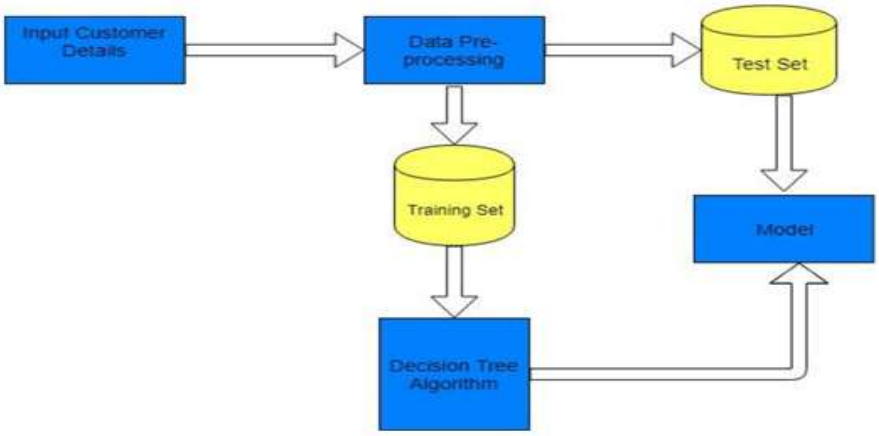

Fig. 4 Architecture Of Proposed Model

Steps For Architecture Of Proposed Model-

1) Input- The Main Highlight Of This Loan Credibility Prediction System Is That It Uses Decision Tree Induction Data Mining Algorithm To Screen/Filter Out The Loan Requests. A Decision Tree Is Developed By Performing Data Mining On An Existing Bank Data Set.
2) Data Pre- Processing –It Is Important Step In Data Mining. Initially The Attributes Are Identified Which Will Help In Making Loan Prediction. Manual Processing Is Also Done. Data Filtering Is Performed After Pre Processing, The Data Set Is Divided Into Training And Test Sets.
3) Decision Tree Algorithm: Decision Tree Is Most Popular Classification Technique. It Is Tree Like Graph. The Decision Tree Algorithm Is Applied To The Training Set. The General Purpose Of Using Decision Tree Is To Create A Training Model Which Can Use To Predict Class Or Value Of Variables By Learning Decision Rules Deduced From Prior Data(Training Data).

V. Conclusion

In This Paper We Have Presented A Loan Credibility Prediction System That Helps The Organization In Making Right Decision. This Project Gives Us Framework For Predicting The Possibility Of Default Of Loan Applicant. This System Is Built Using Data Mining Algorithm.

In This Project The Data Set Contains The Information Of Loan Applicants. CRISP- DM Methodology Is Used To Build The Classification Model. It Consists Of 5 Steps Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation. The Database Is Used For Training And Testing. The Pre- Processing Steps Include Cleaning, Normalization, Feature Extraction Etc. The Data Set Is Trained Using Decision Tree Algorithm. Which Generates A Tree For Prediction Which Successfully Classifies The Customers .Thus This System Is Useful To Predict The Credit Defaulters.

References

[1] Mrs. Bharati M. Ramageri "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal Of Computer Science And Engineering ,Vol No. 4 301-305.
[2] Sudhamathy G., Jothi Venkateswaran C. "Analytics Using R For Predicting Credit Defaulters"2016 IEEE International Conference On Advances In Computer Applications (ICACA).
[3] Https://Www.Investopedia.Com/Terms/C/Creditrisk.Asp
[4] Fabricovoznika ,Leonardo Viana,"DATA MINING CLASSIFICATION".
[5] Dr. K. Chitra , B. Subashini "Data Minig And Its Applications In Banking Sector".
[6] Https://En.Wikipedia.Org/Wiki/Decision_Tree.
[7] Https://Www.Analyticsvidhya.Com/Blog/2016/04/Complete-Tutorial- Tree-Based- Modeling-Scratch- In-Python/
[8] Qasem A. Al-Radaideh ,Eman Al Nagi "Using Data Mining Techniques To Build A Classification Model For Predicting Employees Performance"(IJACSA) International Journal Of Advanced Computer Science And Applications, Vol. 3, No. 2, 2012.
[9] The CRISP-DM Model:The New Blueprint For Data Mining Colin Shearer, JOURNAL Of Data Warehousing, Volume 5, Number 4,Pag.13-22, 2000.