ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 47-51

Study of Data Mining Techniques for Credit Risk Analysis and Risk Analysis by Using Decision Tree Algorithm

Divya Kumawat[1], Aruna Pavate[2], Apeksha Waghmare[3], Suvarna Pansambal[4]
[1,2,3,4]*(Computer Engineering Department, Atharva College of Engineering,India)*

Abstract: *Banks fundamental business model depends on financial intermediation by raising finance and lending (mortgage, real estate, consumer and companies loans). Consumers and companies loan is one the major source of income for banks. Some risks are always involved in lending loan to Customers. This paperlists different data mining algorithms with its uses.It also present a model to predict Credit defaulters as well as loan pay ability of person requesting loan by using data mining techniques. The system uses ID3 algorithm to identify credit defaulters. It also undergoes training to generate the rule set .*

I. Introduction

Data mining analyzes the data from various perspectives and summarizes the data into valuable information. Data mining helps the banks to look for hidden pattern in a group and discover unknown relationship in the data. These data analysis techniques concentrate on extracting quantitative and statistical data characteristics. These techniques provide useful data interpretations for the banking sector to avoid customer attrition. Banks fundamental business model depends on financial intermediation by raising finance and lending (mortgage, real estate, consumer and companies loans). Consumers and companies loan is one the major source of income for banks. In order to reduce the time and efforts taken to verify the customer’s details, banks have started using electronic ways. Risk is inherent part of bank’s business. Granting a loan to customer always involves some risk. Banks has started using credit score for deciding whether to grant loan to the customer or not. Our model also The classifies the loan request based on customer data, credit score and behavioral analysis. Credit scoring decides whether to grant loan to the customer or not. Data Mining’s Decision Tree Induction Algorithm is used for extraction of information from large database. The decision tree algorithm reduces ambiguity in decision-making and provides clarity and transparency in decisions .Hence, the model developed will provide a better credit risk assessment, which will potentially lead to a better allocation of the bank’s capital.

II. Literature Survey

Data mining tools, using large databases, can facilitates in building the predictive model and offers the hidden patterns present in data. It makes the biggest challenge of granting a loan to a customer quite easy. There are many algorithms available for data mining; these algorithms are mainly classified as supervised and unsupervised learning. The concept of supervised & unsupervised learning in data mining is derived from machine learning. Supervised learning tries to clarify the behavior of the target as a function of a set of independent attributes or predictors while unsupervised learning there is no previously-known result to guide the algorithm in building the model. The table 1 and 2 lists some of the supervised learning and unsupervised learning algorithms respectively.

Table I: Data Mining Algorithms for Supervised Functions

Algorithm	Function	Description
Decision Tree	Classification	Decision trees pull out predictive information in the form of human understandable rules. The rules explains the decisions that lead to the prediction[1]
Generalized Linear Models (GLM)	Classification and Regression	GLM implements linear regression for continuous targets and logistic regression for classification of binary targets.
Minimum Description Length (MDL)	Attribute Importance	MDL considers that the best way of explaining the dtat is to have the simplest, most compact representation of data.
Naive Bayes (NB)	Classification	Naive Bayes predicts by using Bayes' Theorem, which derives the probability of a prediction.
Support Vector Machine (SVM)	Classification and Regression	Different versions of SVM use different kernel functions to handle different types of data sets. SVM classification tries to separate the target classes with the major possible margin. SVM regression finds a

		continuous function such that the maximum number of data points lie within an epsilon-wide tube around it

Table II: Unsupervised Data Mining Algorithms

Algorithm	Function	Description
Apriori	Association	Apriori identifies frequent dataset in the database and infers some association rule showing some trends in the data set For example the items that tend to be purchased together and specify their relationship
k-Means	Clustering	k-Means partitions the data into a predetermined number of clusters. The data belongs to a cluster if it has minimum distance from centroid.
Non-Negative Matrix Factorization (NMF)	Feature Extraction	NMF generates new attributes of the original attributes by using linear combinations. The coefficients of the linear combinations are non-negative.
One Class Support Vector Machine	Anomaly Detection	One-class SVM builds a profile of one class and when applies flag cases which are different from built profile. This allows for the detection of different behavior in dataset.

Data mining can help in targeting ‘new’ customers as well as retaining existing customers for products and services. By discovering customer’s previous purchase and earning patterns the bank can retain the existing customers by offering incentives tailored according to customer’s needs. Customer retaining is a big problem for banking sector along with loan request approval [2]. This section discusses the predictive data mining techniques for the loan assignment in banking sector

1) Classification Methods: In this approach, risk levels are categorized in two groups namely risky group and safe group. The customer with past default history are classified in risky group, whereas the rest of all are placed as safe group. One can use any classification algorithm such as Decision Tree and Rule Induction techniques to build models that can predict default risk levels of new loan applications.
2) Value Prediction Methods: This method attempts to predict expected default amounts for new loan applications. It predicts numeric value and so for the same Neural Network and regression techniques which takes numerical data as target variables can be used

Sudhamathy G, Jothi Venkateswaran C. has proposed a framework to effectively identify the Probability of Default of a Bank Loan applicant. They have used data mining functions available in the R package and UCI repository dataset. Initially, the dataset is pre-processed, reduced and made ready to provide efficient predictions. The final model is used for prediction with the test dataset and the experimental results proves the efficiency of the built model. [3].

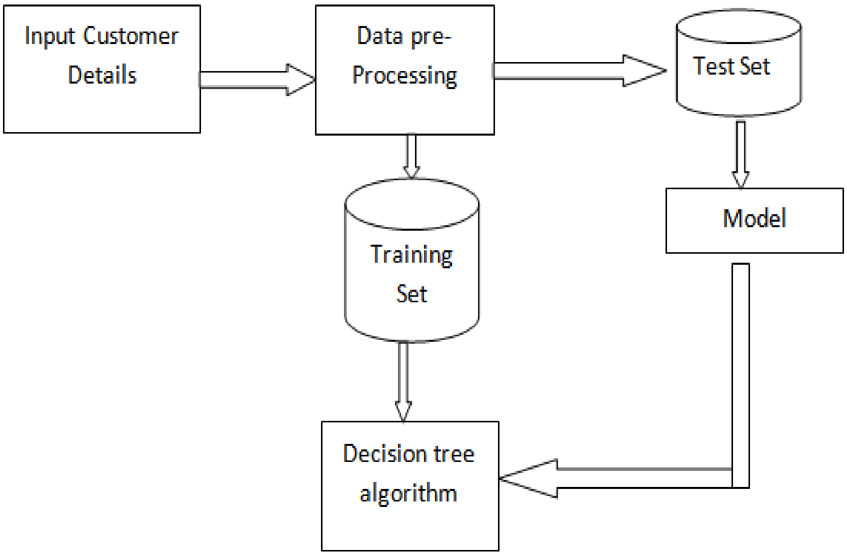
Colin Shearer describes CRISP-DM (CRoss-Industry Standard Process for Data Mining), a non-proprietary, documented, and freely available data mining model. CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These phases help organizations understand the data mining process and provide a road map to follow while planning and carrying out a data mining project[4].

Jingping Chen, Haiwei Pan, Qilong Han , Linghu Chen , Jun Ni have applied, decision tree and information entropy theories to practice of individual housing loan credit risk’s assessment. Under the guidance of the theory of decision tree, we obtain the evaluation of attribute’s importance degree by applying information gain and structuring equation[5].

Mrs. Bharati M. Ramageri, has elaborated the importance data mining in finding the patterns, forecasting, discovery of knowledge in different business domains. Data mining techniques helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry wherever the data is generated. This is the reason for considering data mining as one of the most important frontiers in database and information systems [6].

III. Architecture and working of Proposed System:

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[7].



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.

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)

Test Set: The data used to implement and test this model is taken from the UCI Repository.

Training Set:It is used to tain the classifier

Working of Proposed System:

The methodology followed by system is shown below in fig 2 as follows:

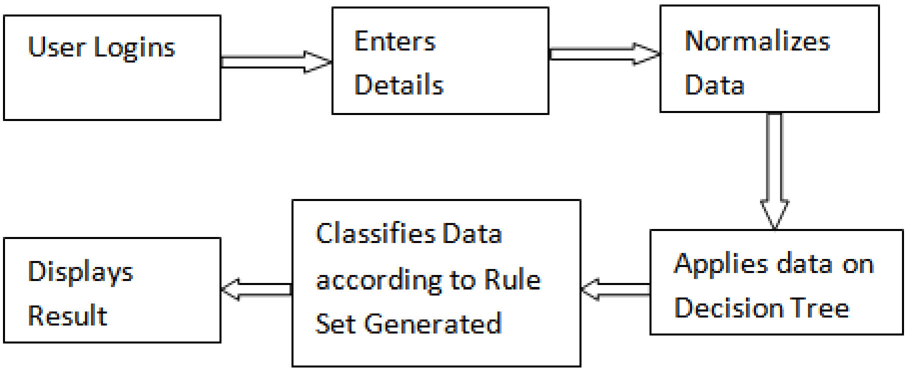

User Logins
Enters Details
Normalizes Data
Applies data on Decision Tree
Classifies Data according to Rule Set Generated
Displays Result

Fig 2: working of proposed system

(1) The system undergoes a training period. First a dataset of 500 values is taken for training purpose. In training the data pre-processing is performed like normalization, data cleaning, removing incomplete data... etc.

(2) After the dataset has undergone training, data mining technique called decision tree algorithm is applied to the dataset.

(3) Data mining is a logical process that is used to search through large amount of data in order to find useful data. This technique is useful in finding patterns in the data which were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of the businesses[6].

(4) Decision tree algorithm develops a rule set according to which a top-down tree will be generated where the nodes will be the attributes. According to which the user will be classified .

(5) The user logins and fills the details in the form after submitting the form the details given is then normalized as shown in fig 3 & Fig 4.

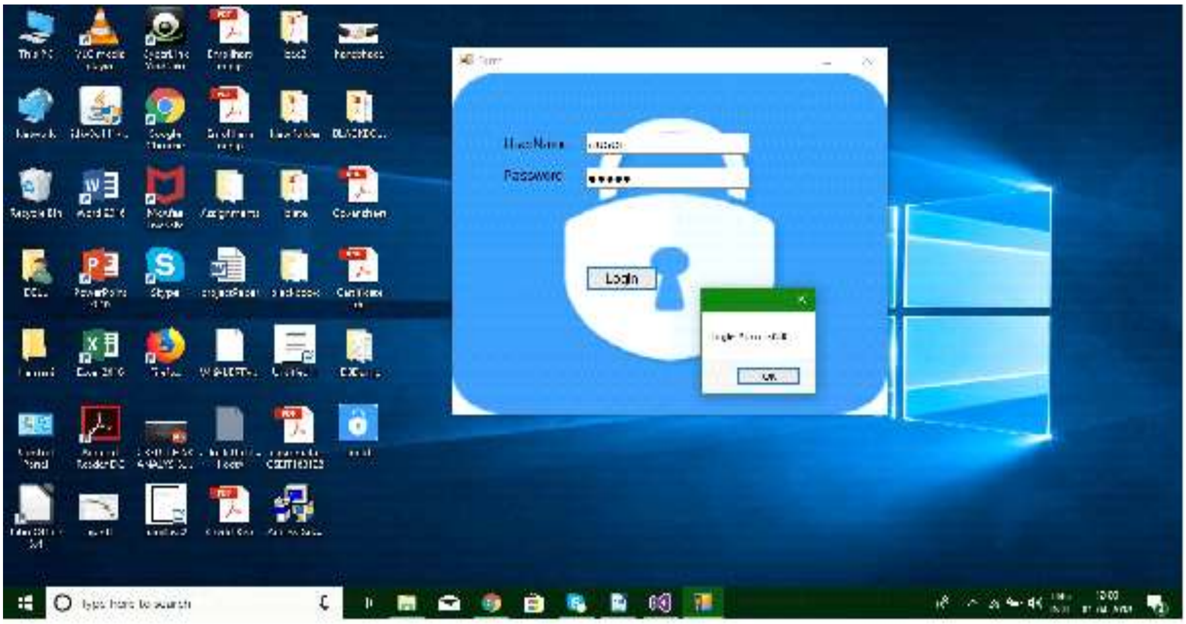
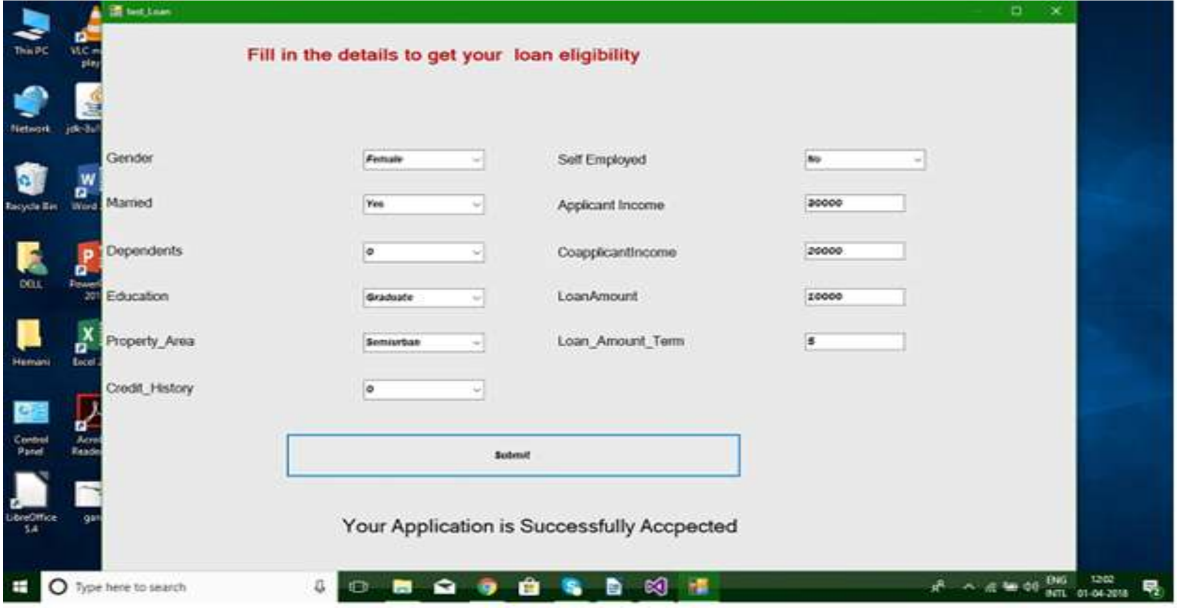


Fig 3: Login window

(6) The decision tree algorithm is applied to it and gives the result according to the rule set generated after clicking on submit button as shown in fig 4**.**



test_Loan

Fill in the details to get your loan eligibility

Gender	Female	Self Employed	No
Married	Yes	Applicant Income	30000
Dependents	0	CoapplicantIncome	20000
Education	Graduate	LoanAmount	10000
Property_Area	Semiurban	Loan_Amount_Term	5
Credit_History	0		

Submit

Your Application is Successfully Accpected

Fig 4: User detail Form along with decision tree algorithm result

IV. Conclusion & Future Scope

The proposed system performs classification without requiring much computation by using decision tree algorithm. It searches whole dataset in order to provide result; It trains the data set by generating understandable rules. The drawback of our system includes non acceptance of null values and requirement of labeled class.

The same system can be built by hybrid techniques of clustering and decision tree method for large dimensional dataset which will show high robustness and generalization capacity. In future the accuracy of data generated can compared with the result generated by other algorithms

References

[1]. Https://Www.Analyticsvidhya.Com/Blog/2016/04/Complete-Tutorial-Tree-Based- Modeling-Scratch- In-Python/

[2]. Dr. K. Chitra , B. Subashini "Data Minig And Its Applications In Banking Sector", International Journal of Emerging Technology and Advanced Engineering , Volume 3, Issue 8, August 2013.

[3]. Sudhamathy G., Jothi Venkateswaran C. "Analytics Using R For Predicting Credit Defaulters"2016 IEEE International Conference on Advances In Computer Applications (ICACA).

[4]. The CRISP-DM Model:The New Blueprint For Data Mining Colin Shearer, JOURNAL Of Data Warehousing, Volume 5, Number 4,Pag.13-22, 2000.

[5]. Jingping Chen, Haiwei Pan, Qilong Han, Linghu Chen, Jun Ni, "Credit Risk Assessment Model Based On Domain Knowledge Constraint", 2008 International Multi-symposiums on Computer and Computational Sciences.

[6]. Mrs. Bharati M. Ramageri "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal Of Computer Science And Engineering ,Vol No. 4 301-305.

[7]. Harshala Yadav, Hemani Yadav, , Divya Kumawat" Credit Risk Analysis" IOSR Journal of Engineering ISSN (e): 2250-3021, ISSN (p): 2278-8719 Volume 11, PP 05-09