IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Application Of FPBDT Algorithm For Decision Analysis On Issuing Vehicle Loan

O.Yamini[1], Research Scholar, Department of Computer Science, S V University, Tirupati
Dr G.V.Ramesh Babu[2], Associate Professor, Department of Computer Science, S V University, Tirupati

Abstract:

More and more data is generated in Internet due to people usage of internet applications for online shopping, insurance has increased leaps and bounds. Which generates huge amounts of which when mined in efficient way will give enormous benefits not only to individuals but also to society. So to tap the potential day by day various methods related to data science and machine learning have been developed. The decision tree classification method is one of the applications of data mining techniques. The decision tree method is a popular and effective technique for classification problems, as it has advantages such as simple algorithm, clear logic, and easy adaptability to different data types. In this paper by overcoming the defects of ID3 decision Tree algorithm development of this application, a new feature probability based decision tree algorithm has been used in the construction of a decision tree model based on the data given, The feature based applications are noted prior to implementation of FPBDT algorithm for generating decision. This can help project managers and decision makers to extract valuable insights from the data and make informed decisions.

Keywords: Decision Tree Algorithm, FPBDT Algorithm, Feature Frequency Table, Feature Selection Table.

1. Decision Tree Introduction

Decision trees are a type of predictive model that use inductive learning based on observation and logical reasoning. They are widely used in various fields for different purposes. A decision tree can represent a system that uses rules to express and categorize a sequence of events that occur sequentially [2].Decision trees belong to the category of supervised machine learning. They are often used because they have several advantages, such as simplicity, interpretability, applicability to different types of data, and reliability [3]. Decision trees begin with a single root. It is a tree that splits into decision nodes and ends with labelled leaves. A basic decision tree is illustrated in As Figure shows that decision trees have roots, branches, and leaves.

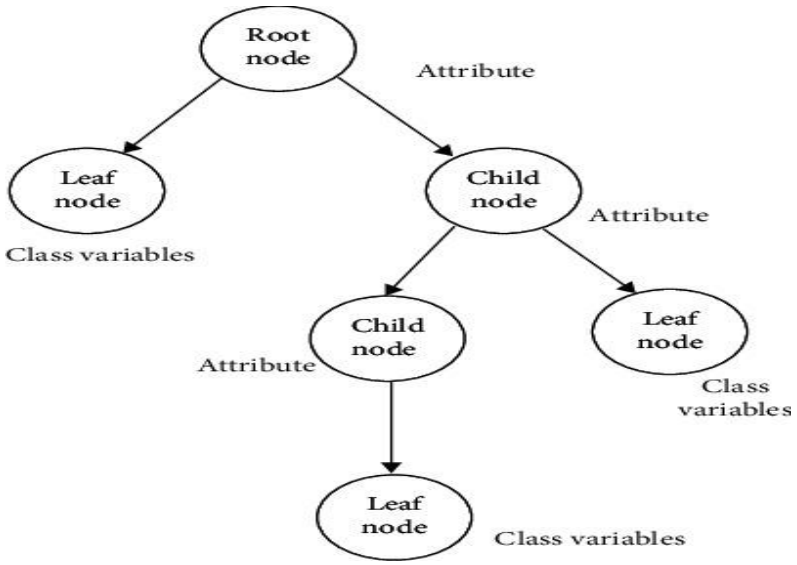

Fig1: Decision Tree

A decision tree is composed of information represented by a tree structure. The tree can be visually displayed by nodes, leaves, and branches. The root node is the starting point of the classification, and it represents the attribute from which all other attributes are derived. The internal nodes or their children have questions about the attribute or the problem [4]. The leaf nodes are the final nodes of the graph, and they represent one of the class variables of the problem. A decision tree is created by two steps: induction of the tree and classification. The initial nodes are built from training sets, and each node has a test attribute and a subset of training data split according to the possible values of that test attribute.

The process of building a decision tree involves splitting the training data into smaller and smaller subsets based on a chosen attribute, and creating a node for each subset. If a node contains instances from more than one class, it is an internal node and the splitting process continues recursively. Various types of Decision trees are given below:

i. ID3 Algorithm

Decision trees are used in various classification learning systems like ID3,C4.5,CART etc., One of the simple decision tree learning systems are ID3 (Iterative Dichotomiser 3) [5], implements a top-down immutable strategy that moves on down and searches only part of the search space. In this process of reaching classification in ID3 entropy and information gain are calculated. ID3 algorithm starts with the dataset as the root node and for every attribute entropy and information gain is calculated. The default entropy used is Shannon entropy in ID3 algorithm. The attribute with the smallest entropy of the largest information gain is chosen for further split [6].

ii. C.4.5 Algorithm

C4.5 is a decision tree based classification algorithm was proposed by J.R. Quinlan in 1993 mainly designed to overcome some of the drawbacks in ID3. Information Gain rate is calculated for test attribute selection [7]. In this algorithm pruning phase in decision tree construction eliminates doubtful branches by swapping them with leaf nodes by backtracking tree.C4.5 will deal with missing values in training set. In C4.5 algorithm, information gain rate is used as the basis of test attribute selection [8]. During construction of decision tree, pruning phase of C4.5 tries to eliminate the un-comfort branches by swapping them with leaf nodes by going back through the tree once it has been generated. The main advantage of C4.5 are it deals with training set with data having missing feature values, deals both discrete and continuous features that support for both pre and post pruning. Algorithm C4.5, in place of information gain a normalized method of “split information is used to overcome the bias in information gain, so we use Split information not information gain.

iii. CART Algorithm

The CART algorithm used in one of the data mining task was proposed by Breiman et al. (1984) is known as Classification and Regression Tree algorithm which creates binary tree with exactly two outcomes from internal nodes. Node splitting is selected by using Gini index. In this algorithm uses binary approach by dividing data set into tow subsets and recursively splits subsets in binary fashion until no longer split is possible. After applying trained data set and tree is pruned, a smallest tree is selected for efficient classification which is the algorithm designed for. CART can be applied with target variable having with both categorical and continuous data, as the tree is known as regression tree if the target variable represents continuous data and otherwise known as classification tree if the target variable contains categorical data, a classification tree can be used. In the CART algorithm, at every root node split rule is applied based on dynamic threshold value entropy is used and whether node to be split or not is computed by using Gini Index. Lower value of Gini index indicates target variable has single category and vice-versa. [9][10][11]

In most of the classification algorithms uses a classifier also known as an algorithm that learns from the training set and then assigns new data point to a particular class. Classification uses mapping function that maps new data entry to class label with help of training dataset which used by mapping function for prediction. Classifications not only maps to single class label but also classifies to more than one. In case of Binary classification there will be two possible outcomes. For example, weather forecast (it will rain or not), spam or fraud detection (predict whether an email is spam or not).In case of Multi-label classification from data set, results in more than two possible Outcomes. For example, classify academic performance of students as excellent or good or average or poor. Classification techniques are also applied in financial markets as part of knowledge discovery for classifying trends of various shares and the automated identification of objects of interest in large image databases.[12][13]

2. Feature Probability Based Decision Tree Algorithm

Feature selection is a process of choosing the most relevant features from a data set to improve the performance of a machine learning model. There are three main types of feature selection methods: filter, wrapper and embedded. Filter methods use statistical measures or information theory to rank the features according to their relevance. They are fast and simple, but they do not consider the interactions between features or the learning algorithm. Wrapper methods use a subset of features and train a model using them. Then they evaluate the model performance using a predefined criterion. They are more accurate and can capture feature dependencies, but they are also more prone to overfitting and computationally expensive. Embedded methods integrate feature selection as part of the learning algorithm. They can optimize both the feature subset and the model parameters simultaneously. They are more efficient than wrapper methods, but they are specific to each learning algorithm. For example, decision trees have an inherent feature selection mechanism [14].

Decision tree algorithm like ID3, C4.5, CART etc., has modified various ways due to disadvantages studied in literature survey like complexity in calculations, memory usage, unable to manage large datasets, decision tree may not be stable in some situations are some of the points considered and designed the proposed Feature Probability Based Decision Tree (FPBDT) Algorithm which uses probability regarding features of data set to decide the in node selection in decision tree. For each table Feature-Frequency-table is constructed to build Feature-Frequency-Table (FFT) with classification (yes/no) and node is selected based on the outcome in the algorithm.

3. Vehicle Loan Dataset Preparation and Transformation for FPBDT

Features of vehicle data are considered and some of rows of Dataset Collected regarding is shown below for Analysis of FPBDT algorithm. Following are some of selected fields shown for processing for vehicle loan purpose.

Loan_ID:- Applicant Id which is given by bank when applied for Loan.

Gender: Male/Female , gender of applicant applying for loan

Married: Marital status of Applicant

Dependents: family dependents of applicant includes children, parents etc.

Education: Qualification of Applicant.(Graduate/NotGraduate)

Self_Employed: Whether Applicant is employed in company or no (doing business)

Applicant_Income: Total annual income of Applicant

Loan_Amount: Loan amount Applied for Loan

Loan_Term: Term indicates number of months of payments of loan instalments.

Cibil : Cibil Score of Applicant applying for Loan.

Property_Area: Property Area of applicant , located in Rural or Urban or Semi-urban.

Loan_Status: indicates the loan can be approved (yes/no).

an_ID	Gender	Married	Dependents	Education	Self Employed	Applican Income	Loan Amount	Loan Term	cibil	Property Area	I St
001002	Male	No	0	Graduate	No	584900	110000	36	700	Urban	
001003	Male	Yes	1	Graduate	No	458300	128000	48	400	Rural	
001005	Male	Yes	0	Graduate	Yes	300000	66000	60	700	Urban	
001006	Male	Yes	0	Not Graduate	No	258300	120000	48	560	Urban	
001008	Male	No	0	Graduate	No	600000	141000	60	650	Urban	
001011	Male	Yes	2	Graduate	Yes	541700	267000	72	457	Urban	
001013	Male	Yes	0	Not Graduate	No	233300	95000	48	600	Urban	
001014	Male	Yes	5	Graduate	No	303600	158000	60	400	Semiurban	
001018	Male	Yes	2	Graduate	No	400600	168000	60	500	Urban	
001020	Male	Yes	1	Graduate	No	1284100	959000	72	400	Semiurban	
001024	Male	Yes	2	Graduate	No	320000	70000	60	780	Urban	

Table 1: Show Applicants data for Vehicle Loan processing.

a) Data Domain Understanding

In any Process domain understanding are essential to get meaningful insights into the model. Before applying any data mining techniques, it is essential to have a clear understanding of the business problem and the domain context. It involves acquiring domain knowledge, such as the characteristics, sources, and quality of the data, the existing solutions or best practices, and the domain-specific terminology. After domain

understanding, data to be described for data mining process, Data description is the process of summarizing the characteristics and features of a data set. It helps to understand the nature and quality of the data before performing any analysis. To describe the data, one needs to consider these aspects: Data source, Data size, Data structure, Data format, Data types, Data relevance.

b) Data Preprocessing Phase

We performed our calculation and simulation on a subset of the entire dataset. This is because the original data had some limitations, such as security related reasons, missing values and unknown values that needed to be addressed and removed before proceeding with the analysis. Therefore, our results are based on a cleaned and processed version of the data.

The primary step involves is Data cleaning is an essential step in data analysis. It involves removing or correcting any errors in the data, filling in missing values, and transforming the data into a format that can be easily analyzed. Missing values refer to data fields that may be empty or include NA, NaN or Null. Depending on the type and extent of missing values, different strategies can be applied to deal with missing values. These include filling in missing values with valid values or dummy values or deleting the rows or columns that contain missing values, if they do not affect the overall results.

c) Data Transformation

Data Transformation for Loan Dataset has to be done for to convert data into a suitable format or categorical basis for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range. All item name which are in string are to be converted / encoded into integers for FPBDT algorithm.

4. Methodology

The KDD Process is a framework for extracting knowledge from data. The KDD process for vehicle loan data can help lenders improve their decision making, optimize their policies and strategies, enhance their customer service and satisfaction, and increase their profitability and competitiveness.

It consists of the following steps:

i)Data selection:choosing the relevant data sources and subsets for the analysis. A data warehouse contains a variety of data, along with other data which is needed to achieve each data mining goal. Financial institution issue loans, bonds, interests etc., for customers and hence large database is maintained. For our research loan data is selected with particular fields essential for loan processing.

ii)Data preprocessing:cleaning, transforming, and integrating the data to make it suitable for mining.KDD process is data cleaning and data preparation, which has to be done before the actual data mining can take place.In Loan Dataset, visualization techniques or in Excel book is used to look for missing data, imperfect data, etc. Or, or data mining tool like knime can be used to look for outliers, which may indicate potentially erroneous data. Statistical methods can also be applied for missing data need.

For our application dataset is analyzed, data transformation is done with values which is essential for our loan processing as each financial organization has set of business rules and data set is normalized to be processed with FPBDT algorithm.

iii)Data mining/Model selection:This step involves applying various techniques to discover patterns, associations, clusters, anomalies, or rules from the data. For example, one may use classification to predict the default risk of a loan applicant, association rule mining to find frequent itemsets of vehicle features and loan conditions, clustering to group similar customers or loans, anomaly detection to identify fraudulent or risky transactions, etc. The proposed algorithm Feature Probability Based Decision Tree (FPBDT) technique is used for filtering and give decision criteria to “approve” loan or not.

iv)Data interpretation and evaluation: This step involves interpreting and evaluating the results of data mining to extract useful knowledge and insights. For example, one may use visualization tools to present the results in an intuitive way, statistical tests to validate the significance of the results, domain knowledge to explain the meaning and implications of the results. Algorithm is implemented in Python and executes on loan data set provides tabular visualsation of FFT and FST for final decision tree creation and selecting customer for loan approval.

v) Knowledge utilization**:** results generated in form of decision tree or rules are stored and applied to both test data to train the model and test data is tested . original data is used by applying FPBDT technique and the extracted knowledge to support decision making, problem solving, or other tasks.

5. Implementation of FPBDT Algorithm for Vehicle Loan Application

The FPBDT algorithm is implemented in Python and Libraries like Pandas, Matplotlib, HTML,statistics etc. ar e used to generate results. Results are generated by executing FPBDT algorithm in with Anaconda spyder and J upyter with Python 3.X. From given Loan Dataset, Algorithm developed into implementation by using Python generated below results. Program generated FFT and FST tables for each Attribute and final prediction of loan approval is given.

```
Feature Frequency Table(FFT) for : Gender
-------------------------------------------------------------------
Gender                Yes      No    Probability
-------------------------------------------------------------------
Male          316    178    0.805
Female        78     42     0.195
-------------------------------------------------------------------
Total Probability   0.6    0.4
-------------------------------------------------------------------

      Feature Selection Table:
-----------------------------------------
Gender            Yes              No
Male          0.64          0.36
Female              0.65          0.35

Feature  :   Gender  is selected  with Feature value  Female ( Female )
having high % of selection 0.65
```

This process has been repeated for all the features in the table to find the root node and its childs to construct a decision tree.

6. Results for FPBDT Technique on Vehicle Loan Data

Decision Tree Nodes are.....

```
Gender is  selected with  Female(Female)

Married is  selected with  No(No)
```

```
Dependents is  selected with  2(2)

Education is  selected with  Graduate(Graduate)

Self_Employed is  selected with  No(No)

Applicant_Income is  selected with  High(2)

Loan_Amount is  selected with  Low(1)

Loan_Term is  selected with  High(2)

cibil is  selected with  Excellent(0)

Property_Area is  selected with  Rural(Rural)
```

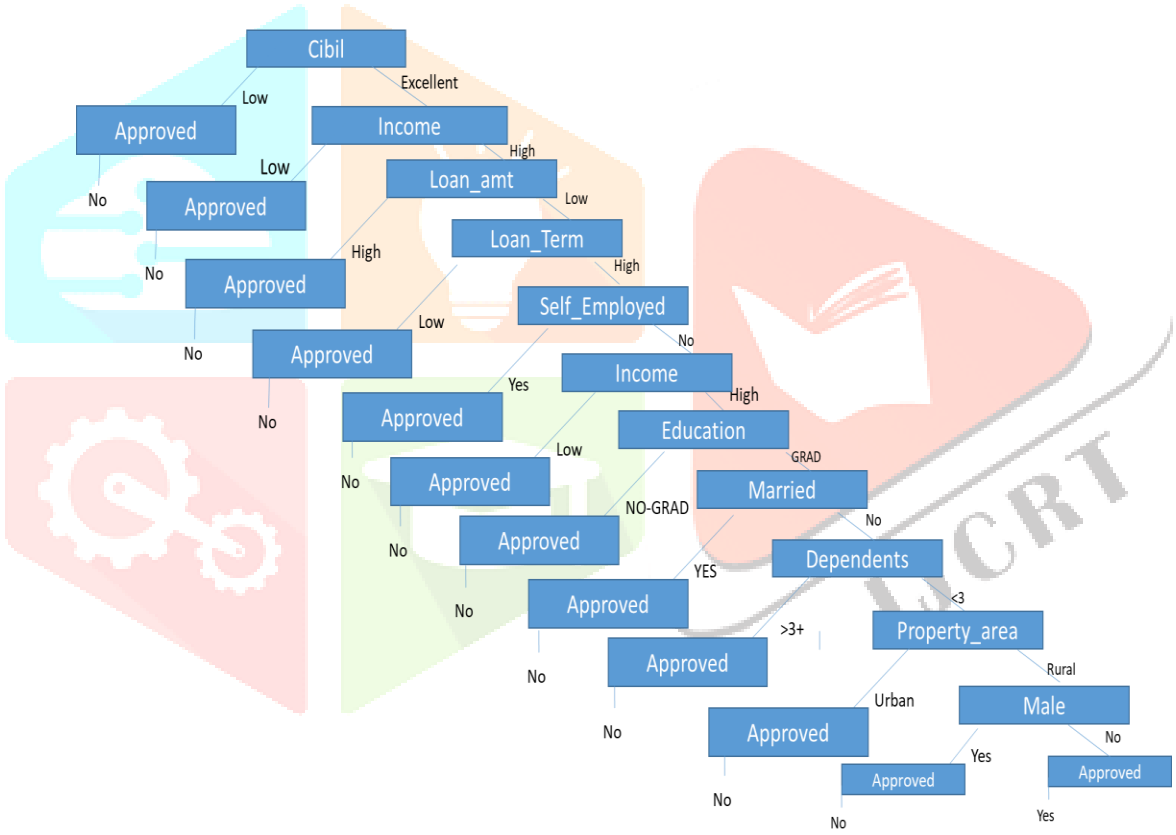

Fig2: Decision Tree Formed with Results Generated above

7. Applications of FPBDT technique

The FPBDT as any other decision tree algorithms can be applied to various fields and domains, such as medicine, education, marketing, finance, etc. Some examples of usage areas are:

- Decision making in Sanction of loans in financial enterprises

- Diagnosing diseases or predicting survival rates based on medical records

- Recommending products or services based on customer preferences or behavior

- Detecting fraud or anomalies based on transaction data

- Classifying text or images based on their content or features

- Predicting house prices or stock returns based on market data

-Prediction of student to continue in same university or not for further study

7. Conclusion

This paper has provided applications related to Data mining and decision trees are given. The feature based applications are noted prior to implementation of FPBDT algorithm for generating decisions. Decision Algorithm Feature Probability Based Decision Tree which is part of research is executed with sample dataset which is transformed and normalized to be suitable for FPBDT algorithm and results are generated showing customer loan approval after giving test data with real time loan data set.

References

[1]. Zhong, X. (2023). The Application of Decision Tree ID3 Algorithm in the Analysis of Enterprise Marketing Strategy. In: Jansen, B.J., Zhou, Q., Ye, J. (eds) Proceedings of the 2nd International Conference on Cognitive Based Information Processing and Applications (CIPA 2022). CIPA 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 156. Springer, Singapore.

[2]. Jui-Sheng Chou, Shu-Chien Hsu, Chih-Wei Lin, Yu-Chen Chang, Classifying Influential Information to Discover Rule Sets for Project Disputes and Possible esolutions, International Journal of Project Management,Volume 34, Issue 8,2016,Pages 1706-1716,ISSN 0263-7863.

[3]. Singh K. The comparison of various decision tree algorithms for data analysis. International Journal Of Engineering And Computer Science . 2017;6(6):21557–21562. doi: 10.18535/ijecs/v6i6.03. [CrossRef] [Google Scholar].

[4]. Kaur H., Wasan S. Empirical study on applications of data mining techniques in healthcare. Journal of Computer Science . 2006;2(2):194–200. doi: 10.3844/jcssp.2006.194.200. [CrossRef] [Google Scholar].

[5]. J.R.Quinlan, "Induction of Decision Tree", Machine Learning Vol -1 ,pp, 81-106, 1986

[6]. A. Rajeshkanna, K. Arunesh, ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

[7]. A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011 IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[8]. Batra, M., Agrawal, R. (2018). Comparative Analysis of Decision Tree Algorithms. In: Panigrahi, B., Hoda, M., Sharma, V., Goel, S. (eds) Nature Inspired Computing. Advances in Intelligent Systems and Computing, vol 652. Springer, Singapore. https://doi.org/10.1007/978-981-10-6747-1_4

[9]. Chien-Liang Lin & Ching-Lung Fan (2019) Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan, Journal of Asian Architecture and Building Engineering, 18:6, 539-553, DOI: 10.1080/13467581.2019.1696203

[10]. B.R. Gains, "Tranforming Rules and Tree into comprehensive knowledge structures" U.M.Fayyad, G.Platetsky-shapiro,p.syth and R.Uthurusamy, eds., Advances in knowledge discovery and data mining , pp205-228, AAAI/MIT Press, 1996.

[11]. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn 16, 235–240 (1994). https://doi.org/10.1007/BF00993309

[12]. Sen, P.C., Hajra, M., Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_11

[13]. Weiss, S. I., and Kulikowski, C. 1991. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Francisco, Calif.: Morgan Kaufmann

[14]. A. P. Bentir, A. H. Ballado and M. J. P. Macawile, "Feature Relevancy Evaluation Based on Entropy Information," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 2018, pp. 1-5, doi: 10.1109/HNICEM.2018.8666381.