A Case Study on Car Evaluation and Prediction: Comparative Analysis using Data Mining Models

Pravarti Jain
Gyan Ganga Institute
Technology &Science Jabalpur, India

Santosh Kr Vishwakarma
Gyan Ganga Institute
Technology & Science Jabalpur, India

ABSTRACT
At the point when an individual consider of buying a car, there are many aspects that could impact his/her choice on which kind of car he/she is interested in. There are different selection criteria for buying a car such as prize, maintenance, comfort, and safety precautions, etc. In this paper, we applied various data mining classification models to the car evaluation dataset. The model created with the training dataset has been evaluated with the standard metrics such as accuracy, precision and recall. Our experimental results show that decision trees are the most suitable kind of dataset for the car evaluation dataset.

Keywords
Data-mining, Text mining, Naïve Bayes algorithm Recommendation system, Car Evaluation data, Rapid Miner

1. INTRODUCTION
Data mining, the extraction of hidden predictive in progression from very big databases, may be an great new technology by means of large potential to support companies thinks on the foremost very important information in their knowledge warehouses. Data processing tools predict future trends and behaviors, permitting businesses to create practical knowledge-driven selections. The automatic, prospective analyses offered by data processing move on the far side the analyses of past events provided by retrospective tools typical of call emotionally supportive systems. Data processing tools will answer business queries that business queries that historically were too time overwhelming to resolve. They clean databases for hidden patterns, predictive information that specialists could miss as a result of it lies outside their expectations.

Understanding the idea in settling on a choice in procuring a car is fundamental to everybody particularly the first buyers through purchasers or any individual who are unpracticed in how the car business functions. Generally we require an car as a methods for transportation however as we include fun into it we overlook the elements that we ought not think little of, which could lead us to relinquishing our wellspring of transportation and again backpedaling to driving, which is not a terrible way but rather includes an excessive number of bothers and is less agreeable as though you have your own particular car that you could utilize right when you require it.

Arranging a decent car from an average to a terrible one are generally being done physically with the assistance of our neighborly mechanics who instructs us to purchase this along these lines or from the conclusion of our family and companions who had past experienced with car inconveniences. It would have been pleasant to have a contraption that can check auto elements and tell that it's an car. In the event that there's such thing then there ought to be no stresses in accomplishing a specific car. In the present time it is dependably the car salesperson identity which urges us to purchase this car or not. We may or won't not know it deliberately but rather we are essentially overlooking the components that would help us fiscally, serenely, and securely in a long run.

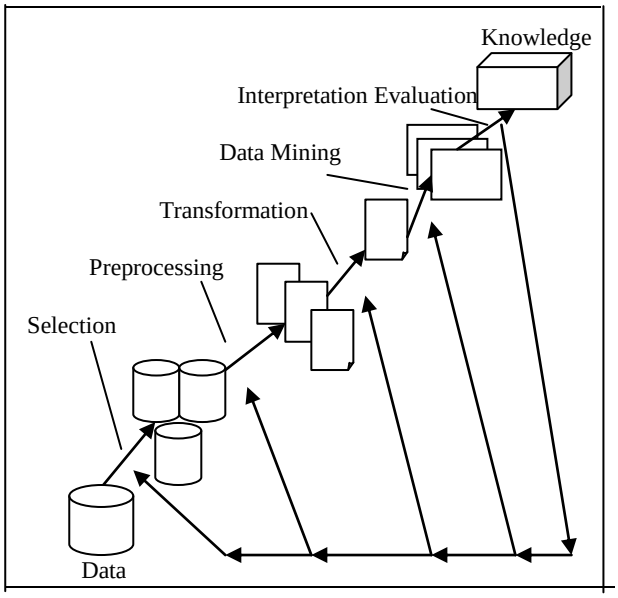

Fig 1: Data mining process

2. RELATED WORK
A study conducted by [1] on the car evaluation dataset employs various data mining technique to investigate the performance of various classifiers. In the research conducted by [2], they also mine customer feedbacks and extract interesting patterns from the dataset and created clusters. The observations as claim by [3] that summarization task is different from traditional text summarization. They proposed a set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods. Their experimental results indicate that the proposed techniques are very promising in performing their tasks. Reviews are not only useful to common shoppers, but also crucial to product manufacturers. In the paper [4], the author proposed an approach for exploring large corpora of textual customer feedback based on labeled clusters in graphical fashion and claim improvement in the accuracy of the applied methods.

Table 1. The model evaluates cars according to the following concept structure

1. PRICE buying maintenance	overall price buying price price of the maintenance
2. TECH	technical characteristics
3. COMFORT door persons luggage boot safety	Comfort number of doors capacity in terms of persons to carry the size of luggage boot estimated safety of the car

Input qualities are imprinted in lowercase. Other than the objective idea (CAR), the model incorporates three moderate ideas: PRICE, TECH, and COMFORT. Each idea is in the first model identified with its lower level relatives.

The Car Evaluation Database contains cases with the auxiliary data evacuated, i.e., specifically relates CAR to the six input attributes: buying, maintenance, doors, persons, luggage boot, and safety.

The Data should look like this:

Number of Instances: 1728 Number of Attributes: 6

Table 2. Attributes Value

buying	v-high, high, med, low
maintenance	v-high, high, med, low
doors	2, 3, 4,5
persons	2,4,5
luggage boot	small, med, med
safety	low, med, high

3. METHODOLOGY

3.1 Rapid Miner

The device which utilized as a part of research is Rapid Miner; It was presented in 2001 by Simon Fischer Ralf Klinkenberg and Ingo Mierswa, and at the Artificial Intelligence Unit of the Technical University of Dortmund. At the outset it is named as YALE (Yet another Learning Environment). In 2006, its headway was driven by Rapid-I, an association set up by Ingo Mierswa and Ralf Klinkenberg .In 2007 the name of hardware has been changed from YALE to Rapid Miner.

According to Bloor Research, Rapid Miner gives 99% of an advanced logical course of action through configuration based frameworks that speed transport and reduce botches by nearly discarding the need to make code. Rapid Miner gives machine learning and information mining frameworks including data pre planning and portrayal, data stacking and change (Extract, change, stack (ETL)), insightful examination and accurate showing, appraisal, and association. Fast mineworker is worked by java programming. Rapid excavator give Graphical UI so here we can arrange and execute sensible work forms. These procedures are called as methodology in fast mineworker and these systems contain numerous operators. This operator has some predefine undertaking with algorithm or coding to perform.

3.2 Naive bayes

Naïve Bayes is a basic strategy for building classifiers models that dole out class names to issue examples, spoken to as vectors of highlight esteems, where the class marks are drawn from some limited set. Guileless bayes classifiers are immediate classifiers that are known for being direct yet particularly beneficial. The probabilistic model of Naive Bayes classifiers relies on upon Bayes' speculation and the elucidating word Naive begins from the supposition that the segments in a dataset are ordinarily free. Before long, the flexibility supposition is consistently harmed, however Naive Bayes classifiers still tend to perform uncommonly well under this farfetched doubt. Especially for little example sizes, Naive Bayes classifiers can beat the all the more extreme choice.

Naïve Bayes classifier is a straightforward probabilistic classifier in view of aaplying Bayes hypothesis (from Bayesian insights) with solid (innocent) freedom suspicions. An innocent bayes classifier expect that the nearness (or nonattendance) of a specific component of a class is random to the nearness (or nonappearance) of some other element. Innocent Bayes classifiers can deal with a subjective number of autonomous factors, regardless of whether consistent or clear cut. Given an arrangement of factors, X={x1, x2, x3… xn}. We need to develop the back likelihood for the occasion Ck among an arrangement of conceivable results C = {c1, c2, c3… ck}. In a more commonplace dialect, X is the indicators and C is the arrangement of straight out levels display in the reliant variable utilizing Bayes' rule

$$P(C|X) = P(X|C)\,P(C)/P(X)$$

$$P(C|X) = P(X_1|C) \times P(X_2|C) \times \ldots\ldots\ldots\ldots \times P(X_n|C) \times P(C)$$

Where to show P (C|X) is Posterior probability and P (X|C) is Likelihood and P(C) class prior probability and P(X) predictor prior probability.

3.3 Decision Tree

A decision tree is a tree-like graph or model. It is more similar to rearranged tree since it has its root at the best and it develops downwards. This representation of the data has the advantage compared with other approaches of being important and easy to interpret. The objective is to make a characterization show that predicts the value of an objective attribute (often called class or label) Based on several input attributes of the Example Set. In Rapid Miner an attribute with name part is predicted by the Decision Tree operator. Every inside node of tree compares to one of the input attributes. The quantity of edges of a nominal inside node is equivalent to the quantity of possible values of the corresponding input attribute. Outgoing edges of numerical attributes are labeled with disjoint ranges. Each leaf node represents a value of the estimation attribute given the estimation of the input attributes represented by the way from the root to the leaf. Decision Trees are produced by recursive partitioning. Recursive apportioning implies more than once part on the values of attributes Fig 2 to shown process decision tree.

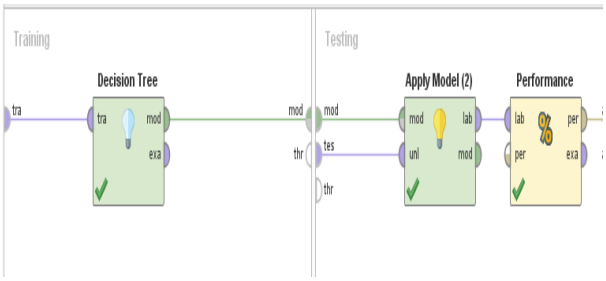

Fig 2: Decision Tree

Fig 3 shows the flow of main process. Process documents from files operator is used for reading text data available in any document file. Validation operator is used for providing training and applying different data mining algorithms in any process.

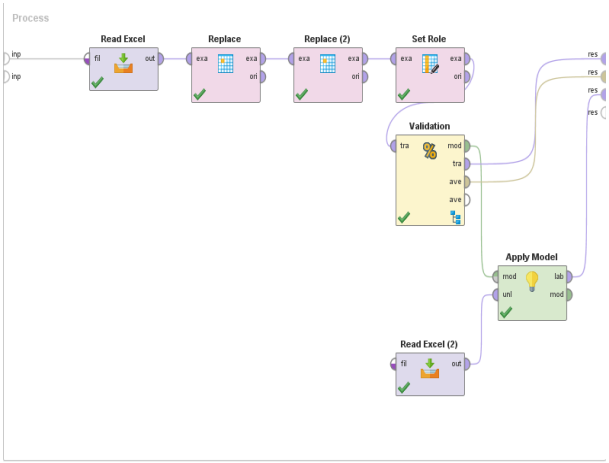

Fig 3: Main process

3.4 Training Dataset

Dataset that is utilized as a part of this work, we prepared them with the content mining administrators accessible in Rapid Miner before applying to the classifiers for preparing, and also testing. For giving preparing, need to gather audit and ordered them physically into 4 types of class names Unacceptability, Acceptability, Good, V-good. These class marks will be utilized to prepare the classifier and afterward in light of this taking in the classifier foresee the name of the testing dataset. Dataset used to excel file document in preparing information. The dataset used Number of Instances 1728 and Number of attributes 6 Attributes values buying, maintenance, doors, persons, luggage boot, and safety.

Class Distribution (number of instances per class)

Table 3. Example of class distribution

Class	N# samples	N[%]
Unacc	1210	70.023%
Acc	384	22.222%
Good	69	3.9930%
V-good	65	3.7620%

The text files provided for the testing are being predicted in one of the predefined label- Unacceptability, Acceptability, Good, V-good using naive bayes classifiers.

3.5 Validation

The validation all operator criterion results accuracy, Classification error, kappa, weighed mean recall or weighed mean precision Fig 4 show the flow of process validation operator only decision tree is enable operator but all operator is disable. Produces a K-NN This operator produces a k-Nearest Neighbor display from the information. This model can be an order or relapse demonstrate contingent upon the information the model works very well on the training set but does not perform well on the validation and Random forest This operator produces an arrangement of a predefined number of arbitrary trees i.e. it creates an arbitrary woodland. the Random Tree operator the model works extremely well on the preparation set however does not perform well on the approval and Naïve bayes This operator generates a Naive Bayes classification demonstrate A Naive Bayes classifier is a basic probabilistic classifier in view of applying Bayes' hypothesis (from Bayesian measurements) with solid (naive) freedom suspicions and rule induction This operator enlistment This operator takes in a pruned set of standards regarding the data pick up the model works very well on the training set but does not perform well on and Decision Tree for characterization of both ostensible and numerical information A choice tree is a tree-like diagram or model Decision tree operator used to testing and decision tree gives the best result all correctly predicted the model works very well on the training set to validation operator table 4 is show below.

Table 4. Criterion result

criterion	K-NN (%)	Random forest (%)	Naïve bayes (%)	Rule induction (%)	Decision tree (%)
accuracy	77.9	81.2	86.4	88.8	91.1
Classification error	22.0	18.7	13.5	11.2	8.8
kappa	0.39	0.51	0.68	0.76	0.80
Weighed mean recall	36.7	37.7	65.2	79.1	78.3
Weighed mean precision	83.7	37.5	78.6	83.6	78.0

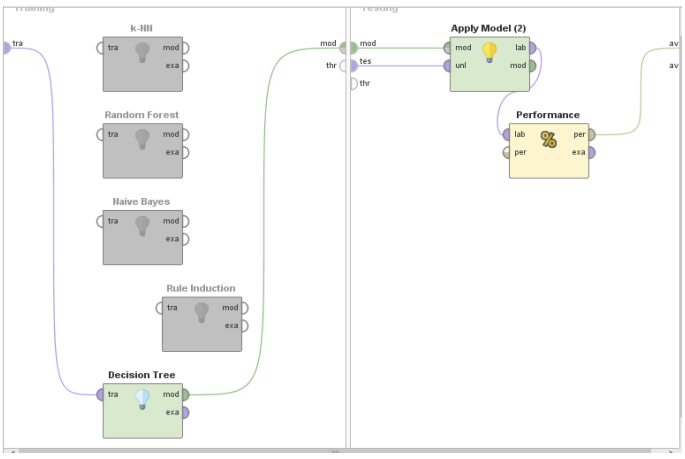

Fig 4: Process validation operator

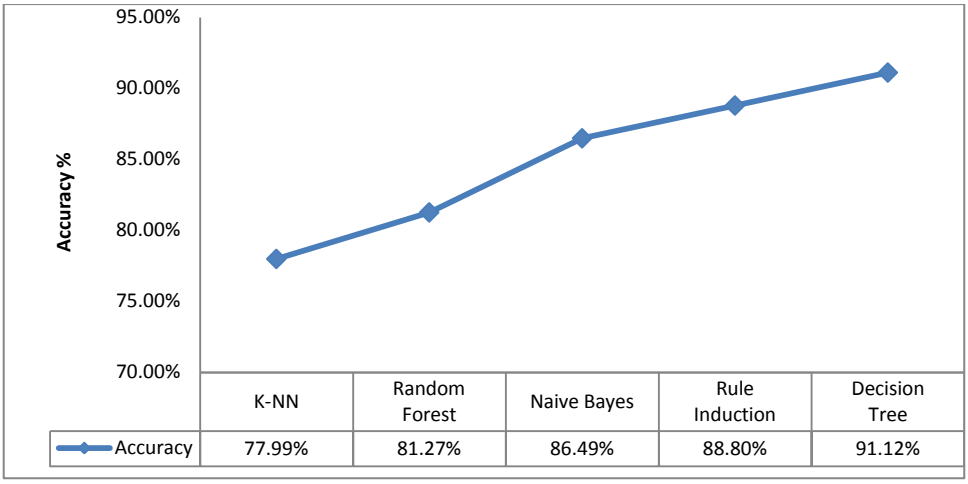

Fig 5: Accuracy graph of decision tree

Fig 5 shows the flow of accuracy graph of decision tree accuracy percentage five operator utilized K-NN, random forest, naïve bayes rule induction, and decision tree. They are all operator give the good result but not correctly predicted and decision tree gives the best result all correctly predicted the model works extremely well on the training set to validation operator then other operator.

4. RESULT

Row No.	result	persons	doors	buying	maint	lug boot	safety
1712	acc	4	5	low	low	small	med
1713	good	4	5	low	low	small	high
1714	unacc	4	5	low	low	med	low
1715	good	4	5	low	low	med	med
1716	vgood	4	5	low	low	med	high
1717	unacc	4	5	low	low	big	low
1718	good	4	5	low	low	big	med
1719	vgood	4	5	low	low	big	high
1720	unacc	5	5	low	low	small	low
1721	acc	5	5	low	low	small	med
1722	good	5	5	low	low	small	high

Fig 6: Sample of dataset

Fig 6 shows the flow of simple of dataset car evaluation dataset used to exceed expectations record archive in planning data The dataset utilized Number of Instances 1728 and Number of characteristics 6 Attributes values buying, maintenance, doors, persons, luggage boot, and safety and qualities esteem the principal traits buying interior v-high, high, med, low and second ascribes maintenance inside to v-high, high, med, low or third credits doors inward to 2, 3, 4, 5 and forward ascribes person inner to 2, 4, 5, and fifth properties luggage boot, inner to small, med, big and six traits safety low, med, high and result 4 types of class names Unacceptability, Acceptability, Good, V-good.

accuracy: 91.12%

	true unacc	true acc	true vgood	true good	class precision
pred. unacc	349	8	0	2	97.21%
pred. acc	17	98	1	5	80.99%
pred. vgood	0	1	16	1	88.89%
pred. good	0	8	3	9	45.00%
class recall	95.36%	85.22%	80.00%	52.94%	

Fig 7: Accuracy of decision tree

Row No.	predicti...	confide...	confide...	confide...	confide...	Buying	maint	Doors	Persons	Luggage	Safety	Result
1	unacc	1	0	0	0	vhigh	vhigh	2	2	small	low	?
2	unacc	1	0	0	0	vhigh	vhigh	2	2	small	med	?
3	acc	0	1	0	0	vhigh	med	2	5	big	med	?
4	acc	0.083	0.917	0	0	vhigh	med	2	5	big	high	?
5	good	0	0.250	0	0.750	low	low	5	5	med	med	?
6	vgood	0	0	0.750	0.250	low	low	5	5	med	high	?

Fig 8: Prediction and confidence classification of decision tree

5. CONCLUSION

In this paper, we evaluated the different classifiers for car evaluation dataset. Based on the customer feedback about the cars used, the model is very appropriate to judge the best car segment as per the requirement of the customer.

In future, research can be use more refine technique to give more accuracy and deal with the some other issue like choose the nature of feeling, also assemble the traverse of the testing dataset and can take a gander at the more auto evolution as enormous number of flexible car are available in market. Not simply with compact brand however for other thing we can perform same investigation.

6. REFERENCES

[1] Ronald F., 2004 "Decision Model for Car Evaluation Final Project in Pattern Recognition.".

[2] Marko B. and Rajkovic V. 1988 "Knowledge acquisition and explanation for multi-attribute decision making." 8th Intl Workshop on Expert Systems and their Applications.

[3] Eduard A. S. and E. K. Özyirmidokuz 2015 has purposed a system named as "Mining Customer Feedback Documents" international journal of knowledge engineering.

[4] Minqing H. B. L. Department of Computer Science from University of Illinois at Chicago 851 South Morgan Street Chicago, IL 60607-7053 "Mining and Summarizing Customer Reviews." American association for artificial intelligence.

[5] Gamon M. and Aue A., S. Corston-Oliver, and Eric R. Natural Language dispensation Microsoft Research, Redmond, WA 98052, USA "Pulse: Mining Customer Opinions from Free Text".

[6] Nicolas C. and Ziegler, Skubacz M. Maximilian Viermetz has work on "Mining and Exploring Unstructured Customer Feedback Data Using Language Models and Tree map Visualizations".

[7] Murali K. P. work on "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining".

[8] Marcelo D. M. and Renate J. S. work on sentimental analysis named as "Using Sentiment Analysis to Assess Customer Satisfaction in an Online Job Search Company".

[9] UCI Machine Learning Group [online] ftp://ftp.ics.uci.edu/bar/machine-learning-databases.

[10] Josef K., and Fabio R. (Eds.) 2000 "Lecture Notes in Computer Science" 1857 'Multiple Classifier Systems', First International Workshop, MCS 2000 Cagliari, Italy.

[11] Duda, R. Hart. Peter E., Stork, David G., "Pattern Recognition" 2ndEdition p. cm. "A Wiley-Interscience Publication." Partial Contents: Part 1. Pattern classification.

[12] Ganjikunta, R. S., "A Study on Multiple Classifier Systems", Computer Science and Engineering, MSU Project for CSE802.

[13] ittler, J., Member, IEEE Computer Society, Mohamad H., Robert P.W. Duin, and Jiri M. "On Combining Classifiers". Centre for Vision, Speech and Signal Processing, School of Electronic Engineering, Information Technology, and Mathematics, University of Surrey, Guildford GU2 5XH, United Kingdom.

[14] Tin K. H., Member, 1994 IEEE, Jonathan J. Hull, Member, IEEE, and Sargur N. Shihari, Senior Member IEEE "Decision Combination in Multiple Classifier Systems". IEEE Transactions on Pattern Analysis and Machine Intelligence.

[15] Zapan B., Bohance M., Demsar J. and Bratko I. work on "Feature transformation by function decomposition" to appear in IEEE.

[16] Bohance M., Zapan B., Bratko I. and Cestnik B. work on "A function-decomposition methed for development of hierarchical multi-attribute decision models".

[17] Zapan B., Bohance M., Bratko I. and Demsar J. work on "Machine learning by function decomposition" to appear in ICML.

[18] Bohance M. and Rajkovic V. work on "Knowledge acquisition and explanation for multi-attribute decision making".