

A Comparative Study of Data Mining Techniques for Preferences of Shopping at the Mall

Dr. M. Senthamil Selvi¹, Mrs. S. Jansi Rani² ¹Prof/IT, ²AP/IT

^{1,2}Sri Ramakrishna Engineering College, Coimbatore-22, Email:jansi.sankar@srec.ac.in.

Article Info Volume 83 Page Number: 8906 - 8911 Publication Issue: March - April 2020

Abstract

Data analysis is a process used to discover useful information from dataset to make conclusions, and to support decision-making. In data mining classification is one of the widely used techniques for categorical data. Past studies indicate that analysis of preferences of Shopping at the mall gives better result using decision tree algorithm (Classification Techniques). The idea of this paper is to conduct the comparative analysis and gives the accuracy measures for classification techniques like decision tree and HP BN Classifier (HP- indicates High Performance) model for preferences of shopping at the mall dataset. This paper models the techniques using SAS Enterprise Miner and concludes which technique is fit for the given dataset. This paper used preferences of shopping at the mall dataset from various people lived in the urban area..

Article History Article Received: 24 July 2019 Revised: 12 September 2019 Accepted: 15 February 2020 Publication: 09 April 2020

Keywords; Classification Methods, Preferences of shopping at the mall dataset, SAS Enterprise Miner, HP BN Classifier

I. INTRODUCTION

In the competitive business [1] world, the preferences of shopping at the mall dataset are very useful to improve the business strategy at the mall in urban areas. Many criteria influence the shopping behavior of the people at the mall. Data mining techniques gives the better insight about the dataset and produces the better knowledge. The knowledge used to change the promotion strategy and discounts of the shops located at the shopping mall in the urban areas.

To summary the dataset Pivot table in excel used to give the summarize of the data. This summary includes average, sums and other statistics. Pivot table groups together the data in meaningful way. This enables to arrange and rearrange statistics in order to get attention to useful information.

SAS enterprise miner gives free access to powerful SAS software for data mining, statistical analysis and forecasting. This paper uses SAS enterprise miner in SAS OnDemand for Academics.

With the SAS OnDemand for Academics, gives same world class analytics software used by government, business and university sites around the world.

II. RELATED WORKS

Research made for credit risk [2] made by using SAS Enterprise miner. The paper modeled the decision tree classification model and logistic regression classification model, neural network classification model, SVM classification model and LibSVM support vector machine. The study analyzes the accuracy by using different credit data and classification tools and classification predicting models. The result of the papers concludes the best model based on the accuracy measures derived from classification model.

Predicting the characteristics of the people living in the South Africa region using logistic regression and



decision tree algorithm [3]. The paper concludes that decision tree classification model works better for the given dataset then logistic regression.

The aim of this paper has three folds:

(1) Information about the dataset using pivot table in Excel 2013

(2) Construct the model based on the dataset using SAS Enterprise Miner

(3) Finally, compare the model and tells the better one for the given dataset using SAS Enterprise Miner

The paper has organized as follows: Section2 describes the literature survey on analysis and comparison made for different dataset. Section3 describes the Pivot table construction and model construction using SAS Enterprise miner, Section4 gives the comparison of the decision tree and HP BN Classifier model using accuracy measures and Section5 tells which model fit for the preferences of shopping at the mall dataset.

III. DATASET:

Data set used in the paper is shopping preference data based on the questionnaire asked from different people in the urban area. Dataset consists of 500 rows and 4 columns. The questionnaire have the following Gender whether male or female, Marital status Married or Single and finally the target class as Preference to shop in Mall (Yes/No).

Table 1: Shopping Mall Dataset (Table shown only first fifteen respondents)

Respondents	Gender	Marital Status	Preference to shop in Mall				
1	Male	Married	No				
2	Female	Married	Yes				
3	Female	Single	yes				
4	Male	Single	yes				
5	Female	Married	yes				
6	Male	Married	No				
7	Female	Married	yes				
8	Male	Single	No				
9	Female	Single	No				
10	Female	Married	yes				
11	Female	Married	No				
12	Male	Married	No				
13	Female	Married	No				
14	Female	Single	yes				
15	Male	Single	yes				

Summary of the dataset:

To analyze and summarize about the dataset, Pivot table from excel have used. Pivot table consists of Rows, columns, values and filters. I considered Gender and Marital status in rows, and in values consider count of Preference to shop in Mall. Fig:1 shows the final result of Pivot table based on the above criterion. Fig:2 gives the graphical representation of the dataset summary.



Fig 1: Pivot Result of Shopping mall dataset





Fig 2: Graphical representation of Shopping mall dataset Pivot result

IV. MODEL CONSTRUCTION USING SAS ENTERPRISE MINER:

In this paper, dataset is analyzed using SAS Enterprise Miner[1], classification algorithms such as decision tree and Naïve Bayes classification methods. In decision tree, implementation based on gini index and entropy, both gives different results and HP BN classifier has different category like Naïve Bayes, Bayesian Network

. Finally the result calculated by comparing the results of decision tree and Naïve bayes classification algorithms.

SAS Enterprise Miner is enterprise data mining software for modeling and assessing the dataset. SAS Enterprise miner has SEMMA model.

Description about SEMMA:

- Sample
- Explore
- Modify
- Model
- Assess

		Restance to Enterprise Mass	
	w materi		
SAT DANSES	w what		
Marita	w Sertan.		
	W Bathan	C. David New Prand - New (of it Specify Fright) Name and Second Sectors ()	
- Sille	w	Tarrentit is progent service and thermities on the 2012 January for the special of their states are then well be within to the resultion.	
§505	- 14 M		
	11000	Marcella Second	
3		At loss beauty	
		Carry Contractory Contractory	
		LORE LORES LARGE	

Fig:3 : New project creation

The above figure shows the Creation of the e Project and named as Shopping mall

Classification:

Classification method used to classify the categorical attribute. Dataset is divided into training and testing data. Classifier model trained by using training dataset and tested against the test dataset. Prediction for numerical attribute. Regression comes under prediction used to predict ordered or continuous values. Types of learning in the dataset is supervised learning because the class label of the sample is known. (i.e.,) Preference of shopping at the mall – Yes/No Decision Tree:

Decision tree construction is top-down recursive, divide and conquer method. For the construction of the decision tree, attribute selection measures used. The measures are gini index and entropy.

HP BN Classifier:

In this, High Performance Bayesian network the following are the categories in SAS Enterprise miner 1. Naïve bayes

2. Bayesian network.

Architectural diagram to compare the classification models:





Fig:4 : Architecture Diagram Model construction:

The above Architectural diagram explains, first create the dataset by collecting the information from the target people. The created dataset have some human made error and machine made error. For that, pre-process the dataset. The dataset is analyzed using pivot table in excel to get the statistical information about the dataset or meaningful information about dataset.

At the second step: Model is constructed with the classification algorithm like decision tree and HP BN classifier.

Third step: Model comparison (Decision tree and HP BN classifier) and analysis of the accuracy measure of the model is anlysed

Fig:5 Model construction using SAS Enterprise miner The figure shows the model construction using SAS Enterprise miner. In the figure, it has the following modules

1. File Import: Module that can read the files from the location (file format: .csv)

2. Graph explore and state explore: Used for data evaluation.



Fig 6: Results of graph explore module

3. Data Partition: This module is used to partition the dataset. Dataset partitioned into training, testing and evaluating.

4. Decision Tree: Selected from model and Ordinal target criterion as Entropy (it has entropy and gini)

5. HP BN Classifier: In high Performance BN classifier, selected naïve bayes for classification

6. Model comparison: In order to compare the classification models (i.e., Decision tree and HP BN classifier).

	8			
	Dilasti Auto Miticale Separantila			
	Te life the finder			
	COARY			
Ande Ande Desire The Depart services	Co. Company			
e Eff. Time Reday				
2414				
	The Day Destination Man			
N all areas hand for LDI LDI				
H Jall, Ant array (pan) from 1.40 1.40	(a) Into histofical legal localizationers to ma a full legal advolutioners to ma as full			
8 30 Return for All 96.00 20.00	2			
d JM_ Stail lepter of heads . M. H	LB Taget broke Perpenty Trail			
	13 Teart Issue American Incention Data Homenian			
	2 F F 6.50 ELD 2 5.00			
E Constitution fails	12 TO B 15.200 6.000 12 3.700			
	22 B TE 2448 2447 38 1587			
Des körftelt hept heiderfolken je die gehalt begin land-tokens is die is het	CA 70 70 80 8.500 D4 1.50			
	23			
i hugh herar heperg ting	3			
Supt Now Honoug Honoug Dat Honoug	0			
B. R. M.M. Mark, N. 1994	3			
1 M B 1107 1308 17 1376	ar the concerns as			
E 20 12.04 6.05 6 7.08				
1 10 10 TAN NON 21 KAN	c) An analysis of the second of the first second s second second sec			
	The second secon			
	the basic basic binner being			
	and append append contact contact			
Data Constitution Mate				
The resource as				
hes his-field hear-between to me to hell heart into between to me to hell				
	2			
i have free have free	14			
I Reprise Reprise Postore Postore	14 American Tree Tabley			
	16			
	30 Sen his-field Super Inside-Performing to the junction for the initial initia ini			
	14			
	if he			
	14 Taulote 4 Taulote Bale d Notein			
8 amment tota laakagt	of logs lice Lts Lts logue (lingue biscorcial holating			
45	8			





The result has calssification table and event classification table. Event classification table has true postive, true negative, falsenegativa and false postive values. In the above dataset postive is preference of shopping at the mall is Yes and negative sample is prefrences of shopping at mall is No.



			Decision 1	Depth	- HP BN Clas	siter]		-	15 Ro 16 17 Tk0	1e 1 077 10	evel evel	Count					
The Statut	0		y					_			1 1			2			(11)
Selected Model	Producessor	Model Node	Model Description	Target Variable	Terpet Label	Selectori Criterion: Trasi: Misclassifica Son Rate	Train: Sum of Frequencies	Train: Misclessifica Son Rate	Train Maximum Absolute Drar	Tran. Sun of Squared Errors	Train Average Squared Error	Train Root Average Squared Enter	Train Divisor for ASZ	Train: Total Degrees of Freedom	Test Sun af Frequencies	Yeat Sun of Vieghts Tores Preps	Test Miscle Sun Ro
1	Tree HPBNC	Tree HPBNC	Decision Tr. HP BN Cla	Preference Preference	Preference Preference	0.20403 0.382872	397 397	0.20403 0.382873	0.864583 0.820413	127.1446	0.160132 0.195074	0.400165 0.442802	794 794	39	7 103 103	204	0.5

Fig: 9: Model comparison results

Fig 9: shows the model comparison results, the results has all accuracy measures of the selected models. In overall modal comparison results, consider only the output. In the output, it has the value of True Positive, True Negative, False Positive and False Negative. By considering the

confusion matrix value (TP, TN, FP, FN) from SAS enterprise miner for shopping mall dataset, Decision tree (Entropy) fits well then HP BN classifier (Naïve Bayes).

V. CONCLUSION:

Fe Statutos														•			
Selected Nodel	Predecessor Node	Nodel Node	Nodel Description	Target Variable	Target Label	Selecton Orteron: Train: Misclassifica ton Rate	Train: Sum of Frequencies	Train: Misclassifica ton Rate	Train: Naximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Test Sun of Frequencies	Test Sun of Weights Times Freqs	Test Wis cless ton Rate
Y	Tree HPBNC	Tree HPBNC	Decision Tr. HP BN Cla	Preference Preference	Preference Preference	0.20403	397 397	0.20403	0.864583 0.820412	127.1446	0.160133	0.400165	794	397	103	205	0.2

Fig 10: Fit Statistics of the selected models



Fig: 11: Score Ranking of the model

- Misclassification Rates of the models 1. Decision Tree: 0.23 2. HP BN Classifier: 0.368
- 2. HP BN Classifier: 0.368

By considering the measures like True Positive, True Negative, False Positive and False Negative, misclassification rate and also based on the score ranking overlay, Decision tree fits the given preference of shopping mall dataset.

VI. ACKNOWLEDGEMENT:

HTTPS://ODAMID.ODA.SAS.COM/SASODACO NTROLCENTER/

REFERENCES

- Hettiarachchi, N. D., Pilapitiya, S. U., Jayasinghe, N. S., Deemantha, H., &Vitharana, S. (2015)." I Shopping: Intelligent Shopping and Predicate
- [2]. Analysis System Using Data Mining". 2015
 IEEE International Conference on Data Science and Data Intensive Systems. doi:10.1109/dsdis.2015.75
- [3]. Shin-Chen Huang, Min-Yuh Day" A comparative study of data mining techniques for credit scoring in banking", 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), Electronic ISBN: 978-1-4799-1050-2
- [4]. Ramona Serban, AndrzejKupraszewicz, and Gongzhu Hu, "Predicting the Characteristics of People Living in the South USA Using



Logistic Regression and Decision Tree", 2011 9th IEEE International Conference on Industrial Informatics, Print ISSN: 1935-4576, Electronic ISSN: 2378-363X