Commonly used Algorithms in Data Science Along with Internal Logics and Implementations through R Programming

T.Srikanth, Sateesh Nagavarapu, K.Umapavankumar, Narahari D

Abstract: ***The terms machine learning, deep learning and data science are buzz words now a days. The usage of these techniques with some technologies like R and Python is most common in the industry and academics. The current work is dealing with the inherent logics existing in the algorithms like Classification, Dimensionality reduction and Recommender systems along with the suitable examples. Some of the applications mentioned here like Facebook, Twitter and LinkedIn to exploit the usage of these algorithms in their daily usage.***

The discussion about online platforms like Amazon, Flipkart are other areas where the recommender systems were most commonly used algorithms. The outcome of the work is the logical things hidden in the usage of the algorithms and the implementation wise which are packages and functions helpful for the implementation of the algorithms.

The belief is the work will be helpful for the researchers and academicians in the context of algorithmic perspective and they can extend the work by contributing their thoughts and views on the same work. Unlike in the normal programming, R/Python simplifies the logic of algorithms so that the lines of code and understanding of the problem is bit simple when compared with general programming languages.

The work explains the mail respondents related to the allocation of the house by the company as a response to their mail by considering Urban, semi-urban and rural areas of the customers, the income range of the customers also observed in the allocation of the house. The implementations are with R by using classification and the corresponding results were published with the explanation of the values found in the implementation.

Keywords: R, Mail respondents, , Classification, rpart, Factor Analysis.

I. INTRODUCTION

The terms data science, machine Learning and Deep learning are all having their own context and meaning.

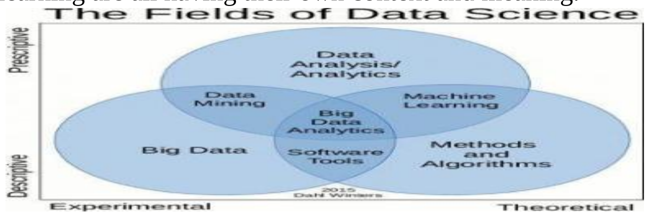

Fig1: The integration of data science with different Concepts.(stoodnt.com)

The reference of Artificial Intelligence (AI) is unavoidable while dealing with data science, Machine Learning (ML) and Deep Learning. AI is the field of study where a machine cab ne added with common sense and make it intelligent. So anything which makes the machine intelligent can be treated as AI.

Machine learning can be treated as branch of AI, where the ML algorithms improve the performance of the given task by learning from the input data. The algorithm learns from the data whenever the input data is getting changed automatically the algorithm follows the change and come up with the intelligent outcome.

ML algorithms might be classified into Supervised learning, Unsupervised learning and Reinforcement learning. Some of the example algorithms were regression, decision tree, random forest and recommender systems. Deep learning is bit advanced to ML in the contextof object tracking, video segmentation which is most frequently used in traffic analysis and image and video captions by just verifying the image or video. Deep learning works based on the Neural network theory which involves working of synopsis and other components.

Data Science is a broader area which involves data pre-processing, applicability of the suitable algorithms and generation of the results. Data Scientist depends on statistical methods which involves predictive modelling, descriptive analytics and prescriptive analytics[1].

In the data science projects there is a flow of activities like a functional problem is converted to statistical model, the statistical model is applied with suitable algorithm, the algorithm is supplied with training data and test data which finally gives the outcome.

The flow of the work is divided into various sections. Section II explains the Classification, dimensionality reduction.Section III explains research issues existing in the social media used ML algorithms. Section IVexplains the conclusion and future scope

II. ALGORITHMS COMMONLY USED IN ML AND THE IMPLEMENTATION

ML algorithms have wide scope in this session the working of classification, dimensionality reduction and recommender system were described.

A.CLASSIFICATION

Classification belongs to the category of supervised learning algorithms, the training data are tagged by labels indicating the class of observations, new data is classified based on the training set.

Revised Manuscript Received on February 05, 2020.

Mr. T.Srikanth, Asst. Professor,CSE, Malla Reddy Inst. of Tech.
Dr.Sateesh Nagavarapu, Associate Professor,CSE, Malla Reddy Inst. of Tech
Dr.K.Umapavankumar, Associate Professor,CSE, Malla Reddy Inst. of Tech.
Mr. NarahariD, Asst. Professor,CSE, Malla Reddy Inst. of Tech

In supervised learning the outcome of the algorithm is clearly specified and input data was with labels[2].

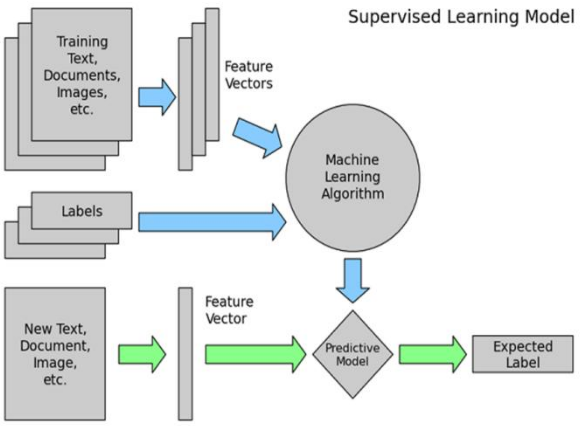
Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Fig2:Supervised Learning Process.

The flow of the supervised learning Model is simply captured using the above representation. The input data might be documents, images from the input data the identification of the Feature Vectors is the initial step. Along with the feature vector the identified labels can be given as input to the suitable ML algorithm.

In the later cases to the implementation model the test data can be given. The outcome of the model can be captured in the form of expected label. Classification algorithm predicts unknown labels. If the data belongs to the categorical data (such as blood group,T-shirt sizes, gender) then the model is known as classification, similarly if the input data is numerical in nature then such prediction belongs to Classification and Regression (CART)[3-4].

CART is preferred because the model gives the better classification and while working with continuous and discrete data in the from of classification and regression.

The implementation of the classification model for the data mail_respond.csv, the model has been created with the packages supported by R language.

```
setwd("C:\\uma")
getwd()
data1<- read.csv("Mail_Respond.csv",head=TRUE)
data1
house<- data1$House.Type
dist<- data1$District
inc<- data1$Income
pc<- data1$Previous_Customer
outc<- data1$Outcome
library(rpart)
library(rpart.plot)
cust_model<-
rpart(outc~dist+house+inc+pc,method='class',control =
rpart.control(minsplit = 10))
cust_model
plotcp(cust_model)
Cust_model<- prune(cust_model,cp=0.03)
Cust_model
rpart.plot(cust_model)
pred<- predict(cust_model,type="class")
pred
```

The above implementation involves the model creation with Mail_Respond.csv data so as to classify the customers based on the kind of the house opted by the customers. The house type label might be detached, semi-detached, Terrace. The dist label belongs to Rural, Urban and Suburban. The inc label holds the data such as the income level of the customer like low or high. The pc label tells whether the customer is previous customer or not. Here the classification label out predicts the allotment of the house to the customers.

The rpart library provided by R having the scope of regression and partition tree which is a classification model. The library rpart.plot having the advantage of displaying the classification model in the pictorial manner.

Thecust_model involves the construction of the model with various labels and prune specifies the elimination of the unnecessary paths in the classification. Here CP stands for Complexity Parameter where the function provides the optimal pruning in the classification.

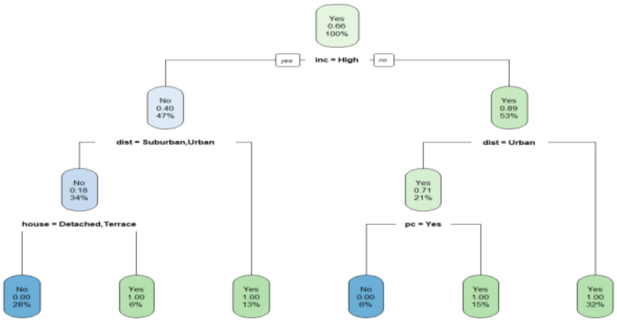

Fig 3: Classification with rpart library of the Mail_Respondant.csv data.

Once the classification is done now the prediction have to perform for the given data, where the model explanation can be observed with summary.

```
Cust_model
********
n= 100
node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root 100 34 Yes (0.3400000 0.6600000)
  2) inc=High 47 19 No (0.5957447 0.4042553)
    4) dist=Suburban,Urban 34 6 No (0.8235294 0.1764706)
      8) house=Detached,Terrace 28 0 No (1.0000000 0.0000000) *
      9) house=Semi-detached 6 0 Yes (0.0000000 1.0000000) *
    5) dist=Rural 13 0 Yes (0.0000000 1.0000000) *
  3) inc=Low 53 6 Yes (0.1132075 0.8867925)
    6) dist=Urban 21 6 Yes (0.2857143 0.7142857)
     12) pc=Yes 6 0 No (1.0000000 0.0000000) *
     13) pc=No 15 0 Yes (0.0000000 1.0000000) *
    7) dist=Rural,Suburban 32 0 Yes (0.0000000 1.0000000) *
```

The above result shows the prediction of the cust_model and the confusion matrix can be displayed for the number of customers who have been allocated the house.

```
pred
outc  No Yes
  No  34   0
  Yes  0  66
```

So 34% of the customer’s outcome label is NO and 66% of the customer’s outcome is YES.

B.FACTOR ANALYSIS

Factor analysis is much commonly used technique in the dimensionality reduction of the data. The purpose is to identify the strongly associated dimensions with larger datasets. The large data set of variables with the usage of factor Analysis can be reduced the complexity which leads to improve the interpretation of complex data[5].

A Researcher want to conduct a survey based on the data recorded related to tooth paste usage by the customers. It involves various factors like shiny teeth and to avoid tooth decay etc. The applicability of the factor analysis is to reduce the number of dimensions which are not that much correlated to the given data within the input.

```
## Factor Analysis
setwd("c:\\uma")
getwd()
data1<-
read.csv("Factor_Analysis_Example.csv",head=TRUE)
data1
data1<- data1[,2:7]
data1
#To compute the z-score for normalization of the data
zdata<- scale(data1)
zdata
#checking the correlation
cor(zdata)
library(psych)
#Check for Sampling
KMO(zdata)
cortest.bartlett(zdata)
#set up a model of PCA
Cust_model<- princomp(zdata)
Cust_model
summary(cust_model)
plot(cust_model)
```

The Factor Analysis follows the process of normalization of the data, and there after the correlation exists in the data, later the checking for the sampling then construct the model with principle component analysis.

```
Call:
princomp(x = zdata)
```

Table 1a:Standard Deviations:

Comp.1	Comp.2	Comp.3
1.6248534	1.4643026	0.6533590

Table 1b:standard Deviations:

Comp.4	Comp.5	Comp.6
0.5743539	0.4201674	0.2869988

Table 2:6 variables and 30 observations.
Importance of components:

Importance of components:	Com p.1	Com p.2	Com p.3	Com p.4	Com p.5	Com p.6
Standard deviation	1.62 4853	1.46 4303	0.65 3359	0.57 4354	0.42 0167	0.28 6999
Proportion of Variance	0.45 5198	0.36 9687	0.07 36	0.05 6876	0.03 0438	0.01 4201
Cumulative Proportion	0.45 5198	0.82 4885	0.89 8484	0.95 5361	0.98 5799	1

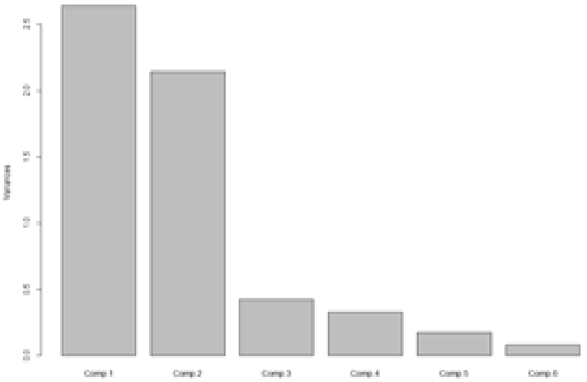

Fig 4:Factor Analysis on the Customer data.

III. RESEARCH ISSUES IN RECOMMENDER SYSTEMS IN FACEBOOK, TWITTER AND AMAZON

In case of FB, Twitter and Amazon frequently uses recommender system algorithms, this section describes the internal logic of the recommender system used in the social media and online shopping websites[7-9].

In case of FB the recommender system suggests the people you may know for the new friend recommendations, in case of the Twitter the news feed you may missed suppose if you could not login to your account from past one week or so forth[10-11].

In case of Amazon the recommendation follows like after selection of the product the same kind of the product with the additional features will be recommended for the customers,which provides the same product with less price. In addition to this the selection of the products and the recommendation system also provides the profile based selection of the products.

The researchers can focus on the fake news and sarcasm which is populated in the social media. The analytics based on the news feeds of the Twitter leads to sentiment analysis requires the positive, negative and fake data. In case of the Amazon the possible improvements are profile based recommendations like based on the user age, qualification and interest areas.

The research objective for the recommender systems involves the following possible research scope.
Identification of the duplicated ID’s of the social media
Identification of the fake ID’s
in twitter, Facebook etc.,

Retrieval Number: C5811029320/2020©BEIESP
DOI: 10.35940/ijeat.C5811.029320

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A procedure to identify the sarcasm kind of the messages so as to differentiate the original news and rumours spread over the social media.

IV. CONCLUSION AND FUTURE SCOPE

The work presented in the paper described the AI, ML and data Science context. The algorithms spread over in the ML like Classification and dimensionality reduction were implemented with R programming and results also published along with the summary of the data.

The classification technique explained in the work with the help of rpart so as to predict the customers who responded to the mails sent by the company. The recommender system related to FB, Twitter and Amazon were described along with the usage of user based profile and content based profiling. The work also exploits the research issues identified in various social media like fake news spreading, duplicated user ID's and sarcasm need to take care, a procedure need to establish to capture all these problems by using the algorithms of classification, clustering and dimensionality reduction related to data posted to the media.

REFERENCES:

1. www. Analyticsvidya.com
2. www.datacamp.com
3. Isabelle Guyon, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 2013 , p.1157-1182.
4. C. P. Chen and C.-Y. Zhang, "Dataintensiveapplications, challenges,techniques and Technologies: A survey onBig Data," Inf. Sci., vol. 275, pp. 314–347,2014.
5. I. Mashal, O. Alsaryrah, and T.-Y. Chung,"Performance evaluation ofrecommendation algorithms on Internet ofThings services," Phys. Stat. Mech. ItsAppl., vol. 451, pp. 646–656, 2016.
6. E. Alpaydin, Introduction to machinelearning (adaptive computation andmachine learning series). The MIT PressCambridge, 2004.
7. Isabelle Guyon, An Introduction to Variable and FeatureSelection, Journal of Machine Learning Research, 2013 ,p.1157-1182.
8. Uma Pavan Kumar Kethavarapu, "Various Computing modelsin Hadoop eco system along withthe perspective of analyticsusing R and Machine learning", Vol. 14 CIC 2016 Special IssueInternational Journal of Computer Science and InformationSecurity (IJCSIS), PP-17-23.
9. Kim Hazelwood, Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective, Facebook, 2017.
10. https://machine arningmastery.com/best-machine-learning-resources-for-getting- started/
11. Uma Pavan Kumar K, Various Issues in Hadoop Distributed File System, Map Reduce and Future ResearchDirections, International Journal of Pure and Applied Mathematics, Volume 120 No. 6 2018, 4441-4451, June 24, 2018.