Data Science to Predict & find Missing Values and the Apps which Steal information from Your Device

ANDREW ANTHONY SATHISH, IV Sem M.Tech in CNE, SIT, Tumkuru,

Dr. K G MANJUNATH, CSE Dept, SIT, Tumkuru, Karnataka.

ABSTRACT: A huge amount of data is created daily through various Media. to understand and usage of these data in today's era is utmost necessary to take vital decision or predicting the system to accuracy is the research work.

Keywords: Missing Values, Deep Learning, Machine Learning, Artificial Intelligence, TikTok , True Caller.

I. INTRODUCTION

The central purpose of this research task is to make data available to the end user with ease by using various methods – such as Artificial Intelligence, Machine Learning, Deep Learning and Data Science.

In today's world, data are produced from various sources, some data are accessible and some data are hidden due to privacy and due to legal factors.

The growth of Big Data is multidimensional and very cumbersome as well. They are not easily available or data is missing or data is not clearly understood due to its data structure. Data format, data scripts and etc. The data depends upon its **Volumes, Varieties, Velocities and Veracities.**

II. LITERATURE REVIEW

Data Mining, Data Warehouse, Cloud Computing actually hold and process large amount of data. To find the meaning or Missing values is foremost important to forecast and predict it.

The Main problem is to co-ordinate between DBMS, SQL, XML, HTML , Excel Sheets and etc. to our required format. The various tools to find the missing values or segregating them or compare them for any predictions. Many research work have been carried out but not all research work is good to predict the future.

1. This paper is mainly focusing to find missing values or data and forecast or predict it to its accuracy.

2. To provide a useful information by using the tools and technologies.

III. RECENT DEVELOPMENTS

Recently the Big Data has accomplished a big gain in various fields like News, Sports, Health, Agriculture, Medicine, Science and Technology, History, Culture, Heritage, Epics, Geography, Tourism, Marketing, Automobiles, Advertisements, Movies, Songs, Entertainments, Education and etc. Internet Web scrapping for searching, Indexing and online transaction.

The list is endless; we need to break them into classified sectors according to its size, growth and its availability legally. The raw data has to be processed to some knowledge as per our requirement.

IV. METHODS

The various methods to find missing values are:-

1. **Do-Nothing :** It works on simple rules or algorithm to find the missing values.

2. **By using Avg Values :** in this method the average is found to fill the missing values. It works for figures and numbers but not for facts. It works for numbers in a column. It has lot of incorrect information with uncertainty.

3. **By using 0 or most frequently used values :** it works well for string type or number type missing values by using the most frequently or recently used values. It may also replace the missing values by 0s.

4. **By using similar features or algorithm such as memory:** the k nearest method is used to search the values missing. It uses similar features to find the missing values based on its proximity and closeness. This method can be better off when compared to the previous three methods but its expensive as it needs lot of memory to store the previous information

5. **By using MICE: -** In this method the missing values are entered multiple times. It is used differently for different type of values ie., Binary values are used.

6. **By using Deep learning: -** In this method a set of libraries are used for deep learning to find the missing values. It is better method when compared to other techniques. The only problem with this is at times it may go slow on processing due to large set of information.

V. Data Analysis &Accessibility :

In today's world the growth of data has been of huge size. It has to be stored in memory for future access. These data are stored in remote machines, servers and many more other secondary and auxiliary memories.

The cost is increasing for such huge storage and should be available when accessed for speedy analysis and process. A delay of a second in procuring the data may be a big loss and a delay of a week time may mean nothing or no effect or no loss at all.

Automation and processing is the new development in machine learning. Clustering of data as per its heading mentioned above namely – Sports, Health, Education and etc.

The design of various algorithms in machine learning is essential / important for analysis and easy accessibility.

Multi Linear Regression [MLR] : it is also called as multiple regression, its a statistical method which is used with several explanatory variables to forecast the conclusion of a response variable. The target of multiple linear regressions [MLR] is to model the linear relationship between the descriptive [independent] variables and answer [dependent] changeable.

Simple Linear vs Multiple Linear Regression :

Linear Regression using Simple method

Y=a0 + a1 * x1

Y = Dependent changeable [DV]

Multiple Linear Regressions

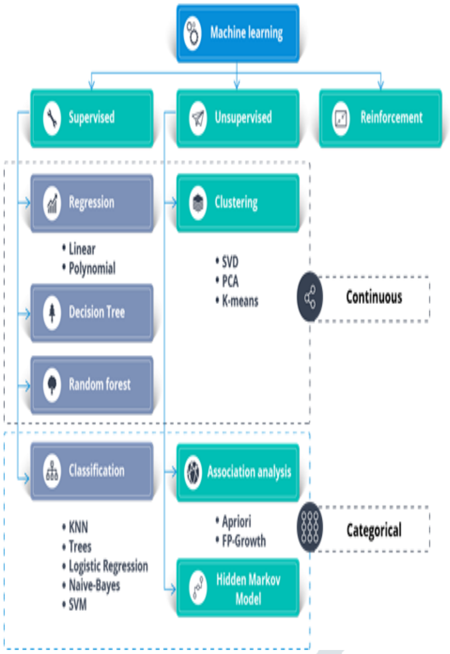
Y= a0 + a1 * x1 + a2 * x2 + … + an * xn

x1, x2, x3, xn = Independent variables [IVs]

Assumptions of a Linear Regression :

1. **Linearity**
2. **Homoscedasticity**
3. **Multivariate Normality**
4. **Independent of Errors**
5. **Short of Multi co linearity**

○ **P value** is a statistical gauge that assists scientists to determine whether or not their hypotheses are correct. P values are used to determine whether the results of their experiment are within the normal range of values for the events being observed.



5 Methods of Constructing Moulds :

1. **All – in**

2. **Backward Elimination**
3. **Forward Elimination**
4. ***Bidirectional Elimination*** ***Step Regression***
5. **Score Comparison**

ALL – In :

Here everything counts.

Backward Elimination

Step 1. choose a significance stage to stay in the mold

Step 2. Fit the complete mould with all achievable Predictors

Step 3. Judge the forecaster with the maximum P-value if P > SL goto step 4 else goto finish.

Step 4. eliminate the forecaster.

Step 5. Fit the mould not including this variable.

Step 6. Finish.

Forward Elimination

Step 1. Choose a meaning level to go into the mould

Step 2. Fit all simple regression moulds y-xn , choose the one with least P –value

Step 3. maintain this variable and fit all possible moulds with one extra forecaster a to the ones we previously had.

Step 4. Consider the forecaster with least P-value, if P value is > SL goto step 3 else goto Finish.

Step 5. Finish.

Bidirectional Elimination

Step 1. Choose a significance level to enter and to stay in the mould. Eg. SLENTER =5%, SLSTAY = 5%

Step 2. execute the next step of Forward Elimination [new values must have P < SLENTER to enter]

Step 3. Perform all steps of backward [old variables must have P < SLSTAY to stay]

Step 4. No novel variables can come in and No Odd variables can exit

All Possible Models

Step 1. Choose a decisive factor of goodness of t = fit

Step 2. Construct all possible regression model 2n-1 total combinations

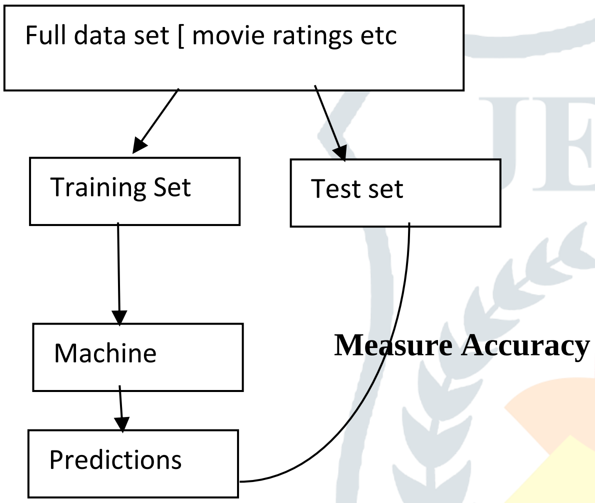
Step 3. Choose the one with the best decisive factor e.g.: 10 columns mean 1023 models.

Evaluating a Recommender systems

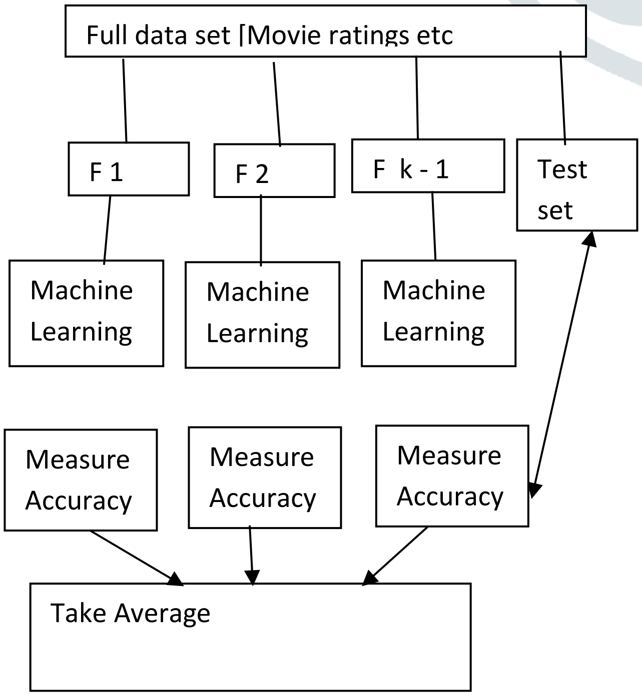
How to evaluate a ML model

- ○ Measure Accuracy of the ML algorithm
- ○ Non live environment
- ○ Live Environment

Train test



K fold Cross Validation



Mean Absolute Error

MAE = Sum of (yi – xi) /n

Predicting Error **yi**	**Actual Rating** **xi**	**Error** **[yi - xi]**
5	**3**	**2**
4	**1**	**3**
5	**4**	**1**
1	**1**	**0**

MAE = (2 + 3 + 1 + 0) /4 = 1.5

Root mean Square Error(RMSE) :

Whole square Sum of (yi – xi)2 / n = RMSE

Predicting Error **yi**	**Actual Rating** **xi**	**(Error)2** **[yi - xi]2**
5	**3**	**4**
4	**1**	**9**
5	**4**	**1**
1	**1**	**0**

RMSE = (4 + 9 + 1 + 0) / 16 = 0 . 87

Top N recommenders :

Hit Rate = Hits / Users

How to Evaluate Hit rate :

- ○ Leave one Out Validation

○ Average Reciprocal Hit rate(ARHR)

○ Collective Hit rate (CHR)

○ Ranking Hit rate (RHR)

Leave one out Validation:

Average Reciprocal Hit Rate (ARHR) :

Fast Text: - this is a technique to compress the data using Huffman coding or Arithmetic coding. This speeds up the coding. The other coding methods include splay coding. Each method is unique in its nature with respect data compression speed, memory space requirement and the cost for compression.

VI. Stealing Information from your Device such as Mobile. Example – True Caller App and TikTok App.

Is TikTok safe – TikTok crosses 1 billion downloads on google play store that mean to say 1/7 of the world population is using TikTok.

The type of content is it safe or not, what social message does it give. All that is later part of it. What is more haunting now is:-

When you download this App and try to record something the App ask for permissions for camera, SD card. you may think its common in most of the App but the problem is something very serious and dangerous. TikTok does not ask your permission but it steals your location and the information and data from your SIM card. When you go to settings you will not realize this but when you select say English songs or English videos yet to your surprise you will see that you have Hindi and kannada songs and videos as well. Hoe does TikTok know that you are an Indian staying in south India .simple it has stolen your Sim data and the location hence it could give south Indian songs and videos. You can understand how dangerous it is and sinister as well. This is unnecessary and overkill, there is no warning also. TikTok uses http server and not HTTP – Secured server.

VII. Conclusions :

There are so many methods to find the missing values. Each method have its own advantage and disadvantages. Hence we can say that there is not any one perfect method but many, what suits for one method may not suit the other method. There are some protocols and rules which decides the decision making. One [the data scientist] should have enough experience and knowledge to decide and use the perfect match to find the missing values.

Authors Profile :

ANDREW ANTHONY SATHISH,
M.Tech [CNE] IV Sem, SIT,
Tumkuru

Dr K G MANJUNATH [P.hd], CSE,
SIT, Tumkuru

Guided By :

1. Asst. Prof Dr. K. G. Manjunath, [P.hd] Dept. of CSE, SIT, Tumkur
2. Maj. P.Arockia Swamy, Dept. Comp Science, SSCASC, Tumkur 572102.
3. Dr. B L Mukundappa, Dept. Computer Science, University College of Science, Tumkur.
4. Dr. B M. Kusuma Kumari, Dept. of Comp Science, University College of Science, Tumkur