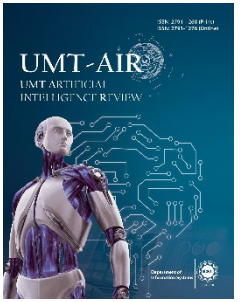

UMT Artificial Intelligence Review (UMT-AIR)
Volume 1 Issue 2, Fall 2021
ISSN(P): 2791-1276 ISSN(E): 2791-1268
Journal DOI: https://doi.org/10.32350/UMT-AIR
Issue DOI: https://doi.org/10.32350/UMT-AIR.0102
Homepage: https://journals.umt.edu.pk/index.php/UMT-AIR

Journal QR Code:

Article: **Automated Exploratory Data Analysis**

Author(s): Hunble Dhillon

Affiliation: Department of Computer Science, University of Haripur, Pakistan

Article QR:

Hunble Dhillon

Citation: D. Hunble, "Automated exploratory data analysis," *UMT Artificial Intelligence Review,* vol. 1, pp. 36–45, 2021. https://doi.org/10.32350/UMT-AIR.0102.04

A publication of the
Dr Hassan Murad School of Management,
University of Management and Technology, Lahore, Pakistan

Automated Exploratory Data Analysis

Hunble Dhillon[1]*

[1]Department of Computer Science, University of Haripur, Pakistan
*Corresponding Author: hubbledhillon@gmail.com

ABSTRACT: This study introduces a novel framework that can be generalized for an automated exploratory data analysis to test a given hypothesis. The current work is about a drug related trend and also provides a specific model to test a motivation-related hypothesis in case of COVID-19. With the utilization of the right, appropriate, and optimized solution available to solve a problem, it is significant that the user feels motivated to delve into the solution for the betterment of the society.

INDEX TERMS: automated exploratory data analysis, motivation-related hypothesis

I. INTRODUCTION

Exploratory Data Analysis is a very important and significant element of determining research direction. These researchers have developed a framework that can be generalized for automated exploratory data analysis to test any hypothesis about a given drug related trend. They have also o provided a specific model to test the motivation-related hypothesis in the case of Covid. With the right kind of solution available to a problem, it is very important that the masses feel motivated to delve into the solution for the betterment of society. Creating right kind of motivation for an already found solution is basically the job of the marketing team of a business house. There exist scientific approaches that attempt to understand the motivation and effect parameters and develop strategies that can lead to a correct solution [1]

Because doing nothing always seems like a safer bet than investing in something. There are some problems that are not mere business issues but are more of survival. Such issues do not have any economic benefits directly tied to them, but are important for survival. That is why scientific methods come into play to help develop the right kind of motivation. Sometimes it is also important to understand what is causing the negative motivation towards the product. For example, one can generate free solar energy but one can find people who are reluctant to avail this opportunity. Neraulink self-driving cars and many more solutions exist. To tackle such situations, it is important to know the real reasons behind the reluctance of people. For example, cost of a new proposed option can be one of the reasons. Polio vaccine is available in Pakistan. But unlike other

countries, it is still not 100% polio-free. Pinpointing the main reason of some negative notions can be very important for drawing positivity towards a product/solution.

To figure out the negative and positive aspects of a motivation-related situation one needs to do multifactor/multidimensional analysis of e data to figure out what is the correlation of different associated factors with respect to the motivation problem at hand. Psychologists have come up with models like the protection motivation theory (PMT) and the theory of planned behavior (TPB) to do statistically analyse data and figure out the relationship of data with different parameters. This is a very common base for starting the analysis of motivation for something like a drug or vaccine [2].

A layman can come up with a number of theories to explain the lack of motivation in general public but only scientific testing can be a sure way to test that hypothesis. PMT is a good explanation utilizing fear appeal [3][1]. It has been tested using the conventional statistics and data set to check those theories. These researchers are going to do something similar by extracting the conventional statistics and utilizing the IAC model to figure out the correlation of different parameters and their effect on the overall Corona virus-related data.
Vehicle automation offers promise for improving safe transportation, access to mobility, and quality of life. However, at least in the early stages of automation, human drivers remain an integral component of the system and their acceptance and use of the automated technology needs to be much better understood [4].
Provide insights into the contributing factors of AV crashes, this study created a unique database from the California Department of Motor Vehicles 124 manufacturer-reported Traffic Collision Reports and was linked with detailed data on roadway and built-environment attributes [5].
This paper reviews approaches for automated pattern spotting and knowledge discovery in spatially referenced data. This is an emerging field which to date has received developmental contributions primarily from researchers in statistics and knowledge discovery in databases (KDD). The field of geographical information systems (GIS) has, however, recognized its

importance as a means for providing more exploratory analysis functionality [6].

Knowledge based system for exploratory statistical analysis of complex systems and environments. Igor has two related goals to help automate the search for interesting patterns in data sets, and to help develop models that capture significant relationship in the data [7].

This paper introduces SmartEDA, which is an R package for performing exploratory data analysis (EDA). EDA is generally the first step that one needs to perform before developing any machine learning or statistical models. The goal of EDA is to help someone perform the initial investigation to know more about the data via descriptive statistics and visualizations. In other words, the objective of EDA is to summarize and explore the data [8].

II. METHODOLOGY

The following method can be used as a framework for any problem that has a bunch of hypotheses and multi-dimension analysis of associated parameters that can help understand the underlying reason for a trend. So in the case of Covid, this research follows the Protection Motivation Theory (PMT) model, which has been the most common explainer of motivation-related questions in many studies about the motivation for a specific drug/vaccine. The researchers have used an extended version of PMT like [1]. But unlike conventional statistics, they have used the well-known model for cognitive computing community interactive activation and competition (IAC) networks. Following is the conventional PMT Model:

A. Protection Motivation Theory

Protection motivation theory was originally created to help clarify fear appeals. (e.g.: appeal for vaccine using death as fear). The theory has the following four major components:-

- **Threat Appraisal / Perceived Severity:** The perceived severity of a threatening event
- **Perceived Vulnerability**: The perceived probability of the occurrence, or vulnerability
- **Coping Appraisal/ Response Efficacy:** The efficacy of the recommended preventive behavior alongside these components.

These researchers have also included the extended prams suggested by the following two because they also went to

access the information source efficacy.

- **Response Cost**: (Which includes money, effort time). Can also be categorized as a coping appraisal?
- **Knowledge**: The accuracy of the knowledge about the vaccine.

The above factors are generic and make very vague sense in terms of Covid. Therefore, the researchers added some concrete factors which could hypothetically answer the motivation questions for them. Following are the parameters used in the current study to build a basis for its hypothesis.

B. Threat appraisal

- Worry of Covid -19 (Scale: 1-10)
- Covid19 vs SARS (Scale: 1-5)
- Possibility of getting Covid in next 1 month (Scale: 1-
- Worry about getting some sort of flu (1-7)
- Worry about getting Covid in the past one week. (scale: 1-7)
- Chances of contracting Covid next month. (1-7)

C. Self-efficacy in having Covid-19 vaccination

- Can you decide about Covid-19 Vaccination? (1-7)

D. Response efficacy of Covid-19 vaccination was assessed using six items

- Believe in vaccination. (1-7)
- Importance for you. (1-7)
- Does Covid reduce risk of Covid-19?
- Vaccination can save lives
- Vaccination is important for health
- Vaccination has positive health benefits

E. Response cost of Covid-19 vaccination 1- Cost

- Time Investment
  - Results

F. Knowledge about the mechanism of Covid-19 vaccination

- Understand how vaccination works
- Understand how a vaccine can help with immunity.
- Don't understand how a vaccine works.

G. Sources of information concerning COVID-19 vaccination

- Internet
- Friends
- Traditional media
- Academic courses
- Medical staff in healthcare institutions
- Coworkers
- Family members

H. Motivation to have COVID-19 vaccination

- Would you get vaccinated?

III. EXPERIMENTATIONS

A. Hypothesis

This paper introduced two hypotheses; Threat Appraisal/Perceived Severity which is the perceived severity of a threatening event like Perceived Vulnerability (the perceived probability of the occurrence, or vulnerability).Coping Appraisal/Response Efficacy (the efficacy of the recommended preventive behavior) Self Efficacy, Knowledge and Response Cost. The two formations are:-

- *Hypothesis1:* Would have a positive relation with motivation to get vaccinated while six would have a negative relationship.
- *Hypothesis2:* Different information sources would have different coping appraisals. It would define relation between information sources and motivation.

B. Measures:

Measures for the above are given earlier in the methodology section. The data set consists of 3145 students. Data was collected from Chinese universities through online surveys where all the above mentioned fields where required.

C. Statistical and Probability Measure

Following four analyses were made to verify the hypothesis, and check their statistical significance. To understand participant demographics following measures were used.

- Mean, Median, Frequencies and T-test and x^2 were used to compare the population with the general population of the Chinese university students.
- Relationships between different parameters were also made using the Pearson correlation, especially to test the hypothesis 2.
- Structural equation modeling was finally used to check the model with the current data of 3145 students.

VI. PROPOSED MODEL RESULTS AND DISCUSSION

To develop a generic model for any kind of given data it was important to do some preprocessing of the data. For that matter, these researchers took the average of all the questions about a given domain/feature and generalized that to (1-5) scale or

to a binary input. So for example, given knowledge about the mechanism, from the three questions given below, we take the average of the given input on a threshold to categories the data on yes no answer[2][4]. By doing so, we can have a direct mapping saying, knowledge+ is in direct proportion to threat appraisal vs knowledge- might have negative. But, we can see the directly associated feature which can really help get the exploratory data analysis. The researchers could not acquire data from the original page and were not worried about the results. They were just trying to create an automated exploratory data analysis framework. So here is some sample data generated through totally fake means using the normal distribution to be more realistic. One has to keep the specific features to have binary values to keep the number of features short but there is no limit to having any number of distinct values for the features.

The statistics for bivariate correlation models to SEM can be simply estimated using the following model. Characteristics analysis for populations is still relevant but once we have a normally distributed population, an IAC model can take care of correlation of different chosen parameters without much mathematical calculation. Until now, the researchers have defined the Modified PMT model for Covid19. Now they will talk about its integration with the IAC Model of the hour. So all the factors defined in the above model can act as features in the IAC Model and jacking up the value for one feature can really tell its association with all the other relevant features like

The price of vaccines and motivation towards getting a shot. The researchers were required to cover the scale given above to some discrete value for features/questions to make more sense in the IAC world. So in Figure 1 in the case of response cost, they had have four different values which were possible values for a response. 1-4 from strongly agree to strongly disagree.

Also, the researchers had not come across any IAC model which was not generated without some human input. But mapping features seemed like a completely robotic task. Therefore, they automated the mapping of feature to hidden layer using a python script that parses the CSV file and maps the relation. For

relation mapping, the logic was simple. When one encounters a new row, one checks what the value for a specific feature is and answers all the values to -1 in the same feature set. From the above circumstances, the researchers have created all the connections, which can be directly used in the PDP Tool based on python implementation. This can run the cycle of specified numbers and update the activation of interrelated features, helping understand the relationship between different features. The result is the higher activation value for any given external input, depicting higher relationship with the external input as in the following figures.

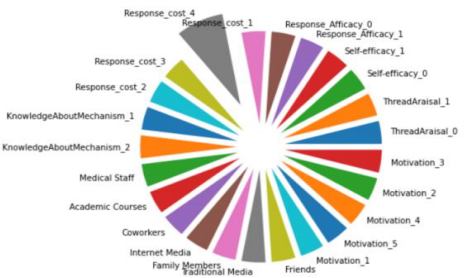

Fig..1.Response Cost has the Highest Input to Model

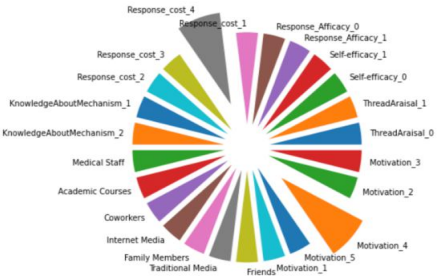

Fig.2. Result of Model after 200 IAC cycles

VI. Conclusion

We conclude that there is a gap between the researches available versus the tools available to add/carry out an exploratory data analysis. Also, some lesser-known cognitive models have their application in different fields of science which may to help make peoples' life easier. Still, there is not enough motivation for research on these avenues. Furthermore, the current researchers see this framework as a first step towards building tools which would make our lives easier in unknown ways.

References

1. P.-W. Wang, D. K. Ahorsu, C.-Y. Lin, I.-H. Chen, C.-F. Yen, Y.-J. Kuo, *et al.*, "Motivation to have COVID-19 vaccination explained using an extended Protection Motivation Theory among university students in China: The role of information

sources," *Vaccines,* vol. 9, p. 380, 2021.
2. D. E. Rumelhart, J. L. McClelland, and P. R. Group, *Parallel distributed processing* vol. 1: IEEE New York, 1988.
3. A. S. Arora, H. Rajput, and R. Changotra, "Current perspective of COVID-19 spread across South Korea: Exploratory data analysis and containment of the pandemic," *Environment, development and sustainability,* vol. 23, pp. 6553-6563, 2021.
4. S. Sreedharan, "Analysing the covid-19 cases in kerala: a visual exploratory data analysis approach," *SN Comprehensive Clinical Medicine,* vol. 2, pp. 1337-1348, 2020.
5. L. J. Molnar, L. H. Ryan, A. K. Pradhan, D. W. Eby, R. M. S. Louis, and J. S. Zakrajsek, "Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving," *Transportation research part F: traffic psychology and behaviour,* vol. 58, pp. 319-328, 2018.
6. M. Boggs, B. Wali, and A. J. Khattak, "Exploratory analysis of automated vehicle crashes in California: A text analytics & hierarchical Bayesian heterogeneity-based approach," *Accident Analysis & Prevention,* vol. 135, p. 105354, 2020.
7. T. Murray and V. Estivill-Castro, "Cluster discovery techniques for exploratory spatial data analysis," *International journal of geographical information science,* vol. 12, pp. 431-443, 1998.
8. R. S. Amant and P. R. Cohen, "Planning representation for automated exploratory data analysis," in *Knowledge-Based Artificial Intelligence Systems in Aerospace and Industry*, 1994, pp. 44-52.
9. S. Putatunda, K. Rama, D. Ubrangala, and R. Kondapalli, "SmartEDA: An R package for automated exploratory data analysis," *arXiv preprint arXiv:1903.04754,* 2019.
10. M. Staniak and P. Biecek, "The landscape of R packages for automated exploratory data analysis," *arXiv preprint arXiv:1904.02101,* 2019.