3-22-2019

An Application of Data Mining of Mental Health Data

Shaikh Shiam Rahman
Georgia Southern University, sr16276@georgiasouthern.edu

AN APPLICATION OF DATA MINING OF MENTAL HEALTH DATA

Shaikh Shiam Rahman
Georgia Southern University
sr16276@georgiasouthern.edu

ABSTRACT

Data mining lies at the interface of statistics, pattern recognition, and machine learning. An organized collection of data and proper data visualization are the main prerequisites of data mining. Proper use of data mining techniques will help identifying important patterns and relationships in a dataset. In this paper, we implement a data mining algorithm on mental health data and find the most important attributes that trigger issues with mental health treatment. For this study we used Microsoft Excel for data preparation and filtering, SQL server as the data storage, and SQL Server Analysis Service (SSAS) for building the data mining model. This is an important process which can help organizations provide a comfortable environment to employees that facing issues with mental health treatment.

Keywords: Mental health, Decision tree, Data Mining

INTRODUCTION

In an environment where mental health concerns permeate nearly every community across the world, there is a responsibility to be frank and real about the issue (Swarbrick 2018). The problem of mental illness amongst working aged individuals and the associated social and economic costs continues to be a major public health challenge for developed countries (Joyce et al., 2016). Tech companies talk about "personal responsibility", "resilience" and "coping techniques" (Swarbrick 2018). But these tech companies do not often talk about the environment that leads to an epidemic mental health issue (Swarbrick 2018). Work stress is associated with psychiatric outcomes of clinical significance that bear great health-care and societal costs (Melchior et al., 2007). Mental health is now recognized as the leading cause of sickness absence and long-term work disability in most developed countries, majority of common mental health conditions are treatable and in some cases preventable (Joyce et al., 2016). It is important to provide an open environment to the employees where they can disclose about health issues.

The purpose of this project is to apply data mining techniques to answer the question whether an employee in a tech company has asked for mental health treatment or not. We aim to identify and disclose very important and useful information that will be beneficial and effective for the tech organizations. This information will be very helpful in making business decisions for the employers and provide a better work ambience to the employees to open themselves in the mental health issue. There are empirically supported interventions that workplaces can utilize to aid in the prevention of common mental illness as well as facilitating the recovery of employees diagnosed with depression and/or anxiety (Joyce et al., 2016).

To identify opportunities for mental health improvement, we plan to identify the most important attributes that cause mental health issues. We will use a data mining algorithm to find the attributes that are responsible for mental health issues. Societal and employer costs associated with workplace mental health constitute a major public health issue (Joyce et al., 2016). This will help the tech companies to identify the facts that need to be considered with high priority to prevent mental health issues.

LITERATURE REVIEW

Decision making is a crucial part of health care. Gnanlet and Gilland (2009) conducts a full-factorial numerical experiment and find that the benefit of using staffing decision under flexibility (demand upgrades and staffing flexibility). The authors discuss 4 configurations under flexibilities. 1) no flexibility, 2) demand upgrades, 3) staffing flexibility, 4) demand upgrades and staffing flexibility. Decision making explores the benefits and trade-offs in employing different types of flexibility. This strategy helps hospital managers to determine optimal staffing and capacity decision making. When capacity decisions are made in the first period and staffing decisions are made in the second period this is considered as decentralized decision making. Where all capacity and staffing decisions are made at the same point in time, this is interpreted as centralized decision making. The experiment concludes that centralized decision making can yield a greater benefit than decentralized decision making.

Using data mining techniques (e.g. association rule, clustering and predictive technique like decision tree) for decision making and establishing relation among the attributes is an old technique. Different data mining algorithms have been used in health care to predict the outcomes or results for future events. Obenshain (2004) mentions three different health care situations where data mining algorithms has been implemented successfully with satisfied outcome. 1) Association rules helped enhancing control over infection. 2) Clustering and Association rules were used to rank the hospitals. 3) American Healthways used

predictive modeling technology to predict the likelihood of short-term health problems. This study compares statistics and data mining techniques, and discusses about their combination to develop the best practice.

The data mining techniques sometimes appear with better prediction performance than dynamic decision making strategies. Meyer et al., (2014) used decision tree to improve the performance of a dynamic decision making system. The authors applied data mining classification techniques to the collected data for discovering the conditions, which were different for dynamic decision-making strategies and use this information to improve the decision-making strategies. This study proposed an iterative approach to determine conditions, under which a given control strategy fails and improve the control strategy using data mining techniques. A predictive data mining technique (decision tree) was used to identify the failed conditions. Different data mining techniques (e.g. decision tree, neural network) were identified that can be used for developing the performances of complex and ill-structured dynamic environment. However, there have many studies which used the direct diagnosis results as decision making attributes, but few studies have used patients' feedbacks as decision making attributes where no direct diagnosis cannot be performed.

Data mining techniques are useful for developing prediction model. Gupta et al., (2018) developed a prediction model for identification of college football players' injury risk. An injury risk assessment model that was developed based on external factors (e.g., exposure to the task, performance role) and intrinsic factors (e.g., injury history, movement efficiency). Logistic regression and Cox regression were used for assessing the injury of football players.

While the importance and efficiency of data mining techniques have been well studied for decision making over direct diagnostics results, very few studies have focused on patients' feedback where diagnosis is restricted. We attempted to fill the gap by using the patients' feedback as decision making attributes.

DATASET DESCRIPTION

Data collection

We have collected this dataset from a survey of 2014, provided by Open Sourcing Mental Illness (OSMI). OSMI measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. OSMI is a non-profit corporation dedicated to raising awareness, educating, and providing resources to support mental wellness in the tech and open source communities. For this survey, there were a total of 1,260 respondents responded from 40 different countries, mainly US based (60% of total data). Most of the respondents are from tech companies. This dataset has a total of 27 attributes and 1,249 successful responses. After analyzing the dataset, correcting the error values and removing the missing/outliers, we have 24 attributes and 1,233 instances for building mining model. Table 1 contains the name, type, description, and sample values of the attributes. The attributes are demographic as well as questions about mental health. We are predicting whether the participant seeks treatment for mental health.

Attribute Name	Attribute Type	Attribute Values	Attribute Description
timestamp	date	8/27/2014 11:29	Time the survey was submitted
age	real	23, 30	Respondent age.
gender	categorical	Male, Female, TransM, TransF	Respondent gender
country	categorical	France, Canada	Respondent country
state	categorical	AZ, AL	only for USA
self employed	categorical	Yes, No, NA	Are you self-employed?
family history	categorical	Yes, No	Do you have a family history of mental illness?
treatment	categorical	Yes, No	Have you sought treatment for a mental health condition?
work_interfere	categorical	NA, Never	If you have a mental health condition, do you feel that it interferes with your work?
no_employees	range value	220-250, 300-350	Number of employees in the organization.
remote_work	categorical	Yes, No	Do you work remotely (outside of an office) at least 50% of the time?
tech_company	categorical	Yes, No	Is your employer primarily a tech company/organization?

benefits	categorical	Yes, No, Do not Know	Does your employer provide mental health benefits?
care_option	categorical	Yes, no, Not Sure	Do you know the options for mental health care provided by your employer?
wellness_program	categorical	Yes, No, Do not Know	Has your employer ever discussed mental health as part of an employee wellness program?
seek_help	categorical	Yes, No, Do not Know	Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity	categorical	Yes, No, Do not Know	Is your anonymity protected if you choose to take advantage of mental health?
leave	categorical	Easy, Very Easy, …	How easy is it for you to take medical leave for a mental health condition?
mental_health_consequence	categorical	Yes, No, Maybe	Would you bring up a mental health issue with a potential employer in an interview?
phys_health_consequence	categorical	Yes, No, Maybe	Would you bring up a physical health issue with a potential employer in an interview?
co-workers	categorical	Yes, No, Some of them	Would you be willing to discuss a mental health issue with your coworkers?
supervisor	categorical	Yes, No, Some of them	Would you be willing to discuss a mental health issue with your direct supervisor(s)
mental_health interview	categorical	Yes, No, Maybe	Would you bring up a mental health issue with a potential employer in an interview?
phys_health_interview	categorical	Yes, No, Maybe	Would you bring up a physical health issue with a potential employer in an interview?
mental_vs_physical	categorical	Yes, No, Do not Know	Do you feel that your employer takes mental health as seriously as physical health?
obs_consequences	categorical	Yes, No	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace
comments	text		Any additional notes or comments

Table 1. Name and Properties of the Attributes

Descriptive analysis

This dataset will provide the details that will help us estimating the mental illness of an employee. In tech companies, the employees are sometimes dealing with extreme pressure and handling other types of mental stress, which may cause mental illness. Here we have all the essential data that can be analyzed to come up with the mental illness condition of an employee.

The target attribute is Treatment (Did an employee ask for mental health treatment) and it has total 615 positive responses (yes) and 618 negative responses (no).

Numerical Data Analysis

Among the 24 attributes we have only one attribute which contains numerical data, age, and the rest of them are categorical.

# of Instances	Min	Max	Lower Quartile	Second Quartile	Upper Quartile	Mean	SD
1,233	18	72	27	31	36	45	7.279

Table 2. Summary of Age

From Table 2, the age of the workers ranges from 18 to 72. The average is about 30 years with standard deviation of 7.279. Table 2 also reveals that 25% of the workers are under 27, 50% are under 30 years old, and 75% of the workers are under 36 years old of age.

Categorical Data Analysis

After the analysis of the dataset, we have identified 2 attributes that can be considered important for predicting whether the employee asked for treatment or not. The most important attributes that we have identified from analysis are: work interference (If they have a mental health condition, do they feel that it interferes with their work?) and family history (do they have any family history of mental health illness). For space constraints, we will not be able to analyze all of the attributes individually. We will analyze two important attributes.

	Seek Treatment		
Work Interference	No	Yes	Grand Total
NA	257	4	261
Never	181	30	211
Often	21	117	138
Rarely	50	120	170
Sometimes	106	347	453
Grand Total	615	618	1,233

Table 3. Summary of Work Interference

How people perform in the work place based on whether they ask for treatment are shown in Table 3. Almost 36% feel that sometimes they have work interference during mental health.

	Seek Treatment		
Family History	No	Yes	Grand Total
No	489	264	753
Yes	126	354	480
Grand Total	615	618	1,233

Table 4. Summary of Family History

Respondents who have a family history of mental health issues is shown in Table 4. According to the Table 4, 61% of respondents do not have a previous family history of mental illness.

All the responses from tech and non-tech organization have been used to build the model without filtering the responses based on the attribute “tech_company”. Majority of the responses are from tech organizations.

METHODOLOGY

As most of the attributes are categorical, the recommended data mining algorithm is decision tree (Larose and Larose 2015) to predict binary outcomes from categorical attributes. For building the decision tree with this dataset, we have used SQL Server Analysis Services (SSAS).

Data partitioning is an important part for preparing the mining model and testing the mining model. We have partitioned our data into two subsets 1) training data and 2) testing data. Training data is used to build the tree. Testing data is used to test the validity of the model. In this dataset we have a total of 1,233 instances. The recommended ratio of training and test instances for given dataset is 20:1 (Larose and Larose 2015). After partitioning, we have a total of 1,172 instances for building the tree and 61 instances to test the model.

SSAS comes with various built-in features, like using different data source for input and output, building the mining model based on different parameters related to that mining model, suggesting the attributes which are best fit to predict the outcome, and choosing the ratio of training and test data as well as tools to validate the model such as classification matrices and lift charts.

Our plan is to build different models based on three SSAS decision tree building parameters: complexity penalty, score method, split method (described below). At the beginning all the models have been built with the parameters' default value. Then we have changed the parameters value to different range and built different models based on those changes.

Complexity Penalty: Complexity penalty inhibits the growth of the decision tree. Decreasing the value increases the likelihood of a split, while increasing the value decreases the likelihood. The default value is based on the number of attributes for a given model: The default value is 0.5 if there are 1 to 9 attributes; the default value is 0.9 if there are 10 to 99 attributes; and the default value is 0.99 if there are 100 or more attributes.

Split Method: Split method specifies the method used to split the node. The available methods are: Binary (1), Complete (2), or Both (3). Default value is 3.

Score Method: Score method specifies the method used to calculate the split score. The available methods are: Entropy (1), Bayesian with K2 Prior (3), or Bayesian Dirichlet Equivalent with Uniform prior (4). Default value is 4.

For each parameter value, we have split the d based on three different attribute sets. As we preferred earlier SSAS has the suggestion feature to suggest the best attributes for predicting the outcomes. We have chosen the first set with attributes suggested by SSAS, second set with attributes which have relevance with the target more than a certain threshold (0.1%) and final set with all attributes.

The parameter value for complexity penalty has been varied among (0.5, 0.9 and 0.99), the parameter value for split method has been varied among (1, 2 and 3) and the parameter value for score method has been varied among (1, 3 and 4).

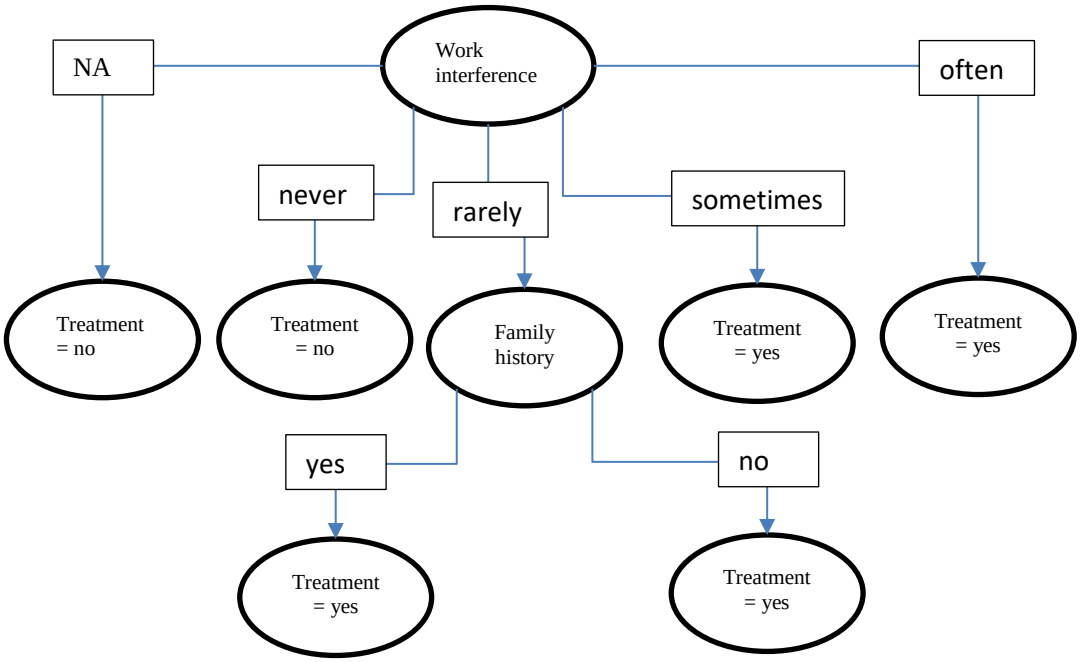

Figure 1. SSAS Decision Tree with Highest Accuracy

We have built total 24 models based on three parameters and different attribute sets of SSAS decision tree building algorithm.

RESULTS

Figure 1 is the decision tree build with complexity penalty value 0.9 and others with default values. Based on the analysis of all 24 models, this was the best model based on test set accuracy. The tree can be expressed as a set of rules.

Rules:

- if work interference = NA, then seeks treatment = no
- if work interference = never, then seeks treatment = no

- if work interference = rarely and family history = yes, then seeks treatment = yes
- if work interference = rarely and family history = no, then seeks treatment = yes
- if work interference = sometimes, then seeks treatment = yes
- if work interference = often, then seeks treatment = yes

77.73% training instances have been identified correctly.

The 3rd and 4th rule, where "work interference = rarely", the Family history is not affecting the outcome. For both the rule "seek treatment = yes" is dominating. 3rd rule, for "Family history = yes" out of 75, 9 is incorrect. 4th rule, "Family history = no", out of 87, 43 is incorrect. The accuracy for 3rd rule is good but the accuracy for 4th rule is not satisfactory.

From Figure 1, it can be seen that work interference and the family history are the most important attributes for predicting the mental health treatment. Table 5 is the classification matrix for decision tree model in Figure 1, here, we can see that, among 61 test cases a total of 53 cases were identified correctly with 8 incorrect prediction. This has 88.5% accuracy.

Predicted	Yes(Actual)	No(Actual)
Yes	35	6
No	2	18

Table 5. Classification Matrix of the Model with Highest Accuracy

CONCLUSION

The objective of this project was to develop a detective system that can assist tech companies to help them identify employees, whether they request treatment for their mental illness. This will help the tech companies to identify whether their employees will ask for the mental health support. 24 different decision trees have been used with different parameter values. The work interference and the family history were counted as the most important attributes for predicting the mental health illness.

From the model, it is clear that the employees with most work interference has the highest probability to ask for mental health treatment. The family history also plays an important role to provoke the mental health illness and ask for treatment. So, the human resource department can prepare a better plan for monitoring the work interference and help their employees.

REFERENCES

1. Gnanlet, A., and Gilland, W. G. (2009) Sequential and Simultaneous Decision Making for Optimizing Health Care Resource Flexibilities, *Decision Sciences*, 40,2, 295-326.
2. Gupta, A., Wilkerson, G. B., Sharda, R., and Colston, M. A. (2018) Who Is More Injury-Prone? Prediction and Assessment of Injury Risk, *Decision Sciences*.
3. Joyce, S., Modini, M., Christensen, H., Mykletun, A., Bryant, R., Mitchell, P. B., and Harvey, S. B. (2016) Workplace Interventions for Common Mental Disorders: A Systematic Meta-Review, *Psychological medicine*, 46,4, 683-697.
4. Larose, D. T., and Larose, C. D. (2015) *Data Mining and Predictive Analytics*. John Wiley & Sons.
5. Melchior, M., Caspi, A., Milne, B. J., Danese, A., Poulton, R., and Moffitt, T. E. (2007) Work Stress Precipitates Depression and Anxiety in Young, Working Women and Men, *Psychological medicine*, 37,8, 1119-1129.
6. Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., and O'Connor, P. J. (2014) A Machine Learning Approach to Improving Dynamic Decision Making, *Information Systems Research*, 25,2, 239-263.
7. Obenshain, M. K. (2004) Application of Data Mining Techniques to Healthcare Data, *Infection Control & Hospital Epidemiology*, 25,8, 690-695.
8. Swarbrick, C. (2018) There's a Responsibility to Be Frank and Real About Mental Health, *The New Zealand Herald*.