Targeted Wellness Benefits Promotion

Maximizing the impact of benefits programs on member health

Problem Statement

- Health and Welfare benefits in the United States are complicated and confusing.
- Employees often underutilize their benefits, costing both employers and employees.
- A personalized approach to promoting benefits, driven by specific behaviors and correlated health outcomes, will help guide employees in navigating and using their benefits.
- Helping employees use the benefits available to them will both help employees and employers save money. It would also help employees better maintain their health and wellness.

Executive Summary

- A study was conducted using the CDC's Behavioral Risk Factor Surveillance System dataset, data collected from a large phone survey of risk behaviors, health conditions, and use of preventative services.
- The study focused on factors related to depressive disorders, and found a number of correlative data elements that may be used to predict depressive disorders.
- In the benefits administration space, we can use this data to help employees of our clients take advantage of mental health related benefits. This will also help our clients maximize their return of investment related to these services.

Related Work

There are a number of studies that have been conducted using the BRFSS data:

Correlative Analysis

- A 2023 study published in the Journal of Public Health and Emergency used the 2020 BRFSS dataset to investigate the association between adverse childhood experiences (ACEs) and self-reported mental health conditions in adults [1]
- A 2019 study published in Preventing Chronic Disease used BRFSS data to assess the prevalence of subjective cognitive decline in adults aged 45 years or older in 49 states [2]

Machine Learning Algorithms

- A 2023 study published in Patterns used the 2021 BRFSS dataset to investigate machine learning algorithms and data augmentation techniques for predicting chronic kidney disease [3]

Data Source

Behavioral Risk Factor Surveillance System (BRFSS)

- The 2022 BRFSS data contains 445,132 records. Each record contains the question results for one survey performed.
- The data has 326 columns, one column per potential question asked during the survey. This includes:
  - Behavioral questions such as "Have you smoked at least 100 cigarettes in your entire life?", "At what kind of place did you get your last flu shot or vaccine?"
  - Health outcome questions such as "Has a doctor, nurse, or other health professional ever told you that you tested positive for COVID 19?", "In general, how satisfied are you with your life?"

https://www.cdc.gov/brfss/index.html

Proposed Work

- The CDC collects a large amount of data nationally, gathering behaviors and health outcomes that we will use in this study.
- This data can be used to find correlations and associations between behaviors and health outcomes.
- By mining these relationships, we can discover questions that we can pose to employees. These questions can help us tailor benefits recommendations to employees directly.
- In addition, we can drive employees to use the benefits they have available to them to positively impact their health and wellness.

Proposed Work

Specifically, I have employed three primary data mining techniques:

- Clustering
  I attempted to use clustering to attempt to identify groups of survey results that are likely to lead to specific health outcomes.
- Association Rule Mining
  I used methods like FP-Growth and Association Rule Mining to help identify survey answers that most likely result in a specific health outcome.
- Classification
  I built a decision tree model to classify specific health outcomes. Decision trees will provide both a classification method, but also explainable parameters we can use to determine specific health outcomes.

Proposed Implementation

- Using the relationships we identify, we can build models to help promote specific benefits that best meet employees' health needs.
- Using a personalized approach to meet employees where they're at, increasing employee utilization of their benefits will have a positive impact on employee's health and wellbeing.
- In addition by providing this as a service, we can improve the return on investment for employers. In addition to benefiting employers, this can help distinguish our company in the market.

Evaluation

To evaluate each of the specific methods above:

- For clustering, I intend on using silhouette coefficients to analyze how well the clustering methods perform.
- For association rule mining, I used support, confidence, and lift to identify the most relevant rules.
- To measure the overall success of the project, success will be determined by if we can accurately determine health outcomes and associated health and welfare benefits using behavioral questions. Implementation will also be important. Determining whether the methods can be reasonably implemented in a real-world setting is critical.

Timeline

The project will be time boxed to an 8 week period total. The project will be divided into the following phases:

- Phase 1 (1 week): Data acquisition, basic data ingestion, and beginning EDA
- Phase 2 (1 week): Extended EDA. Beginning clustering methods.
- Phase 3 (1 week): Beginning Association Rule Mining.
- Phase 4 (1 week): Beginning Classification.
- Phase 5 (2 weeks): Wrapping up all methods.
- Phase 6 (1 week): Conclusion of all work
- Phase 7 (1 week): Final report

Results

- **Clustering**
  - Attempted Hierarchical Clustering across the entire dataset using Gower's Distance as a distance metric. This attempt failed due to complexity of task and limited compute resources.
  - To accommodate above, move to mini-batch K-Means using reduced data via MCA. Ran 150 different K-Means models with varying number of MCA components and Ks. Used silhouette score from each to find the most performant model.
  - Found that Clustering is likely an inappropriate method for this data, because the complexity of the data doesn't yield interpretable clusters.

Results

- **Frequent Pattern Analysis**
  - Began running FP Growth and Associate Rule mining on full dataset. This approach was too memory intensive for resources available.
  - Focused my analysis on a specific health outcome: respondents who answered the question “Have you ever been told you had a depressive disorder (including depression, major depression, dysthymia, or minor depression)” with “Yes”.

Results

- **Frequent Pattern Analysis**
  - *Chronic Conditions and Mental Health:*
    Respondents who have chronic conditions such as arthritis or chronic respiratory diseases are more likely to have a depressive disorder than the overall population. Multiple chronic conditions increase this chance
  - *Disability and Mental Health:*
    Respondents who reported they had difficulty concentrating and remembering were more likely to report depressive disorders.

Results

- **Frequent Pattern Analysis**
  - *Sexual Orientation, Gender Identity, and Mental Health:*
    - Overall, men and women who identify as straight had lower occurrences of depressive disorders.
    - Gay men and women both had higher depressive disorders compared to the overall population (1.54 and 1.88 times the overall population respectively.
    - Women who identify as bisexual are 2.75 times more likely to have depressive disorders when compared to the overall population.
    - Transgender and gender non-conforming populations by far have the highest levels of depressive disorders (2.42 times the overall population for male-to-female respondents and 2.25 female-to-male)

Results

- **Classification**
  - A decision tree was used to discover interactions between various features leading to a "Yes" answer to the depressive disorders question. I then used learned rules to identify interesting relationships learned.
  - "Difficulty Concentrating or Remembering" was the most deterministic feature. A majority of the paths leading to a "Yes" depressive disorder label begin with a "Yes" to this, regardless of other factors.
  - "Difficulty Concentrating or Remembering" answered as "Yes" and "How often have you felt various kinds of stress (specific kinds specified in the question)?" answered as "Always" or "Usually", plus the absence of some other features predicts a 'Yes' answer to depressive disorders with 92.41% and 82.79% probability respectively
  - "Difficulty Concentrating or Remembering" answered as "Yes" and "Difficulty Doing Errands Alone" answered as "Usually", plus the absence of some other features predicts a 'Yes' answer to depressive disorders with 82.79% probability.

Conclusion

Summary

In this study, I focused on how various other health and behavioral factors influence the presence of a depressive disorder. I found that a number of factors were highly related to a respondent having a depressive disorder. In the benefits administration industry, we can use this as a way to help users of our product find appropriate resources that may help them, while also realizing a return on investment for our clients.

Conclusion

Future Work

There are a number of way this work can be expanded on.

- Dive deeper and use more comprehensive data analysis tools to determine indicating factors for depressive disorders.
- More broadly look at combinations of data across multiple categories to find more complex connections.
- Expand data collection to our benefits administration product, collecting data on who uses benefits such as virtual therapy vendors like Talkspace and how that relates to their demographics or behavioral responses.