Caregiver Health - Predictive Modeling

Elena I. Adame

4 June 2022

**Introduction**

This project addresses the health issues of home caregivers in the United States. Many times, caregivers are unprepared for the burden of care they take on and as a result, their health suffers. Medicaid often does not cover the cost of the level of care needed by many caregiving recipients. While the program has worked to expand home based care coverage, the coverage is often optional for states. For states not opted in, this leaves many caregivers in a bind, stuck with covering medical costs of their caregiving recipient as well as providing for their own families. This results in higher levels of burnout in caregivers and negatively impacts their health. This in-turn taxes the medical system as care must then be provided to the caregiver. In 2015, approximately 34.2 million Americans provided unpaid care to an adult family member. (Family Caregiver Alliance, 2016).

As the Baby Boomer generation ages, the need for addressing home health caregiving must be answered. The second largest generation is heading into a state of needing care provided on a regular basis as they age. If the proper care system is not set up, this generation and the others after them will hurt. While the need for support to caregivers was re-addressed at the national level in the 2021 American Rescue Plan, further aid must be mapped out. This project is pertinent to both federal and state governments for them to understand what the current lack of sufficient aid is doing to their citizens. This project aims to show the worsening health of caregivers the longer their care progresses to establish the need for caregiver support, not just to aid the caregiving recipient, but to ensure the caregiver is provided sufficient aid.

The data I used for this project came from the Behavioral Risk Factor Surveillance System, a state-based telephone survey. This data was specifically pulled from the 2015-year group as this survey contained the largest number of respondents who indicated they were caregivers. This dataset contained information on several topics relative to the respondent, to include the respondent’s economic status, demographic, eating and exercise habits, and chronic health information.

**Exploratory Data Analysis**

I began my exploratory data analysis by dropping a large number of columns from the dataset. These columns I identified as not pertaining to my hypothesis. The data dropped in this step related to information about the respondent’s method of taking the BRFSS survey (Section 0.32: Cell Phone Introduction), their access to health care (Section 3.4: Health Care Access), their use of Tobacco (Section 8.4: Tobacco Use), and eating and exercise habits (Sections 11.2: Exercise (Physical Activity) and 10.6: Fruits & Vegetables) to name a few. I deemed this information unnecessary and not useful to the model that would later be built.

As I set out to prove that the health for caregivers was became worse the longer they provided care, I thought it useful to know how their overall health compared to that of non-caregivers. To do this, I established the survey variable, GENHLTH, or General Health, as my initial target variable. This survey question asked respondents to provide an assessment of their overall health in the last 30 days ranging from Excellent (1) to Poor (5). The complete range of responses to this question can be found in Table 2.1.

**Table 2.1: Available Responses to BRFSS General Health Question**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 7 | 9 |
| Excellent | Very Good | Good | Fair | Poor | Don’t Know/Not Sure | Refused |

I separated caregiver and non-caregiver respondents and broke them out further by their response to the GENHLTH question. This can be seen below in Figure 2.1.

**Figure 2.1: General Health for Caregivers and Non-Caregivers**

Chart, bar chart

Description automatically generated

From this figure, I could see that the distribution of data for caregivers and non-caregivers was relatively the same. This did cause some doubt as to the validity of my hypothesis given that the initial trends I was seeing were very similar between caregivers and non-caregivers.

Next, I decided to examine the physical health of the respondents broken out by caregivers and non-caregivers. This relationship can be seen in Figure 2.2 below. The specific question (PHYSHLTH) being examined below is the number of days in which the felt physically unwell over the past month.

**Figure 2.2: Physical Health for Caregivers and Non-Caregivers**

Chart

Description automatically generated

Rather than doing a count plot of the data like Figure 2.1, I chose to show the density of the data for the respondents’ answer to the PHYSHLTH question. The largest area of response for this question was between 0 and 10 days, with another large spike hitting right at 30 days. I include this graph only to show that once again, there are similar trends appearing for caregivers and non-caregivers.

Next, I looked at the physical health response of just respondents who indicated they were caregivers. For this examination, I broke out the responses according to their longevity of providing care (CRGVLNG1). This can be found in Figure 2.3.

**Table 2.3: Physical Health as Related to Longevity of Caregiving (Days to Years)**

Chart, line chart

Description automatically generated

From this figure, I believed my I was on the correct course of proving my hypothesis. I initially thought that the response to the question of physical health showed that caregivers providing care for 6 months to less than 2 years reported significantly higher days of feeling physically unwell. However, what this graph shows is that there is a higher number of respondents who have cared for their caregiving recipient for 6 months to less than 2 years.

**Data Preparation**

I began my data cleansing process by stripping the data of anything unrelated to Section 5.2: Chronic Health Conditions. There were exceptions to this rule. I included information from Section 7.1: Demographics, such as respondent gender, employment status, annual income, and activity limitations, to provide a fuller understanding of the caregivers studied in this project. Finally, in addition to the specific demographic data and the chronic health condition data, I included data from Section 4.1: Caregiver. The information I pulled from this section pertained only to whether a respondent was a caregiver, the length of their caregiving status (months to years), the time of the caregiving status (hours per week), and the services they identified as lacking for them. These four responses were chosen as I felt they related the best to my hypothesis that the length of caregiving worsens an individual’s health.

After narrowing down the topics that would be included in the dataset for this project, I began with the standard process for cleaning the data. As this data pertained to each individual’s specific health care problems or situations, I did not feel comfortable filling in missing data with the mode or median value of each column. As such, I dropped any row in the dataset that had a missing value. There was only one column in which I did replace missing values and that was the Income variable column.

It was during my data preparation that I chose to only have one single target variable. Initially, I had wanted to have the Physical Health (PHYSHLTH) and Mental Health (MENTHLTH) questions as secondary and tertiary target variables. However, as the responses to these questions were numerical rather than categorical data, I chose to include them in the feature data.

Looking at the data, I decided to break out the categorical columns into dummy variables. This was done to ensure that my categorical data would be correctly used when I later built my model. After this was done, I split my data into the Target data (GENHLTH) and Feature data (everything else). From this step, I moved into my model building.

**Model Building**

From working with the data through the exploratory data analysis and cleaning, I determined that because the GENHLTH question had, in total, 7 different responses, I need to work with a multi-class classification model. As such, I did not use Logistic Regression, Linear Regression, or other types of binary classification models.

The first model I chose to build was the K-Nearest Neighbor (KNN) Model. I chose this model as my primary model as I knew that KNN was useful for multiclass classification problems. Additionally, I knew that KNN was good to use when using dataset of low dimensionality. When I determined the shape of my final, cleaned dataset, it was 23,717 x 149. This means that there were 149 features and 23,717 points of data. Because the ratio was low, I felt that KNN would be suitable for this problem set. What I discovered from this model was that it was only 42% accurate. Understanding that this is an arbitrary metric by which to measure a model, I also ran a classification report for the KNN model. The output of this report can be found in Table 3.1 below.

**Table 4.1: Classification Report for K-Nearest Neighbor Model**

|  |  |  |  |
| --- | --- | --- | --- |
| GENHLTH | Precision | Recall | F1 - Score |
| Excellent | 0.34 | 0.16 | 0.22 |
| Very Good | 0.41 | 0.65 | 0.50 |
| Good | 0.43 | 0.39 | 0.41 |
| Fair | 0.43 | 0.30 | 0.35 |
| Poor | 0.54 | 0.25 | 0.34 |

Knowing that for a good classifier, Precision and Recall must close to or equal to 1, I found that this model was not a good model. Precision was below 0.5 for all responses to the GENHLTH question with the exception ‘Poor’. For Recall, only ‘Very Good’ was above 0.5. This model was especially poor at correctly identifying respondents who indicated ‘Excellent’ as it had a Precision of 0.34 and Recall of 0.16. One distinction I noticed for this model is that it produced the highest F1 – Score at 50%. The KNN model also produced the highest Recall value for this project, with ‘Very Good’ at 0.65. For every respondent that answered Very Good, the model was able to predict correctly about 65% of the time.

Given the results for the KNN model, I next built out a Decision Tree model. Decision Tree models are good with complex data. My thought when building this model was perhaps the dataset was to complex for the KNN model with 149 features. Decision Trees are also good at handling numerical and non-numerical data, and while I had created the dummy variables, I thought that this may be an instance of where the Decision Tree will produce a better output. However, this model was only 36% accurate. The classification report results in Table 3.2 show that the model performed poorly for each of the categories of General Health.

**Table 4.2: Classification Report for Decision Tree Model**

|  |  |  |  |
| --- | --- | --- | --- |
| GENHLTH | Precision | Recall | F1 - Score |
| Excellent | 0.26 | 0.27 | 0.26 |
| Very Good | 0.38 | 0.39 | 0.38 |
| Good | 0.40 | 0.40 | 0.40 |
| Fair | 0.34 | 0.31 | 0.33 |
| Poor | 0.36 | 0.34 | 0.35 |

I found that the model could not reach over 50% for any of the General Health categories; this model also shows that the ‘Excellent’ category was still the worst category with regard to prediction.

The final model I built for this project was the Random Forest Model. I chose this model as Random Forest models can handle large datasets. I thought that perhaps my dataset was too large and too chaotic for both KNN and Decision Tree. However, the Random Forest model produced an accuracy of 42%, the same score for the KNN model. Running a classification report on this model confirmed trends that I had seen for both the KNN and Decision Tree models.

**Table 4.3: Classification Report for Random Forest Model**

|  |  |  |  |
| --- | --- | --- | --- |
| GENHLTH | Precision | Recall | F1 - Score |
| Excellent | 0.34 | 0.19 | 0.24 |
| Very Good | 0.41 | 0.49 | 0.45 |
| Good | 0.42 | 0.52 | 0.47 |
| Fair | 0.44 | 0.33 | 0.38 |
| Poor | 0.55 | 0.26 | 0.35 |

This model also provided poor classification for the ‘Excellent’ category of the GENHLTH question. Of the respondents who indicated ‘Excellent’ health, the model was only able to correctly predict this 19% of the time. Additionally, the values for the ‘Poor’ category were very close to those provided by the KNN model.

**Conclusion**

For each of the three models built in this project, the accuracy of the models was determined to be less than 50% accurate. As an accuracy score of 100% is perfect, the models each predicted correctly less than half of the time. This initial metric led me to conclude that my hypothesis was incorrect: Caregivers’ health is not dependent on the length of care they provide to their caregiving recipient. Additionally, it does not appear that the health of caregivers is worse than those of non-caregivers given the trends I observed during my EDA. However, my inability to prove my hypothesis may be because the BRFSS 2015 Survey GENHLTH question relies on a subjective assessment made on the part of the respondent. An article published in Sage Journals (Dunning, Heath, & Suls, 2004) reported that individuals often have an unrealistic optimism about their health. This fact is supported by four different respondents (51, 90, 107, 134) who reported their General Health as 'Fair' or 'Good' despite reporting feeling physically or mentally unwell for at least 15 days or more from the past month. One of these respondents reported feeling both physically and mentally unwell for 30 days. From the classification reports, the models were least precise and sensitive for the ‘Excellent’ category in GENHLTH. I believe this is due to respondents once again being overly optimistic about their general health. Further analysis of the respondents who indicated ‘Excellent’ is needed. The KNN and Random Forest models were more precise for the ‘Good’, ‘Fair’, and ‘Poor’ categories. I believe these categories potentially contain respondents who were more understanding and realistic of their health.

At this time, these models are not ready to be deployed. Further action is required to clean up the dataset to produce more accurate results. I suggest combing through the existing data to identify those individuals that indicated having several of the chronic health issues and examining their response to the general health question. It is my recommendation that any respondent that indicates they have 3 or more chronic health issues and reports they are ‘Excellent’ or ‘Very Good’ be removed from the study to remove unrealistic viewpoints that could skew the data. Additionally, I would also remove the chronic health conditions such as Cancer or Skin Cancer. In hindsight, those two responses should have been removed from the initial dataset as they are not directly related to caregiving, and I feel they bear no weight in this project.

Further challenges also include proving that a caregiver’s health is worse than a non-caregivers health. For this project, I operated under the assumption that a caregiver’s health was worse than a non-caregiver, however, I have not proved this. The initial EDA performed on this dataset showed that the trends between the two categories were very similar when it came to general and physical health. From what was found in this project, this fact needs to be proved or disproved. I believe taking an equivalent, random sample of respondents who are caregivers and non-caregivers could help me reach this result.

While the three models produced results that mirror a random guess, I still believe that support is necessary for caregivers. While this project focused on the general health of an individual as it relates to caregiving, it may be better to direct the focus on the mental and physical health of the individual. While I strayed from making these my target variables as they would produce a classification model with 30 different classes, I believe that respondents were more truthful in answering these questions than they were the general health question. Having personally been a caregiver for an elderly relative, I know the toll it takes on an individual. However, what I failed to consider is that perhaps the toll was primarily on my mental health versus my overall health. Refocusing the project and providing federal and state governments with an understanding of the impacts to mental health would also help these entities learn how to better support their citizens through resources and specified funds.

# References

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace. Psychological Science in the Public Interest. *Sage Journals*, 69–106.

Family Caregiver Alliance. (2016). *Caregiver Statistics: Demographics*. Retrieved from Family Caregiver Alliance: https://www.caregiver.org/resource/caregiver-statistics-demographics/