Project 9

Proposal

Kristie Kooken

Due: April 14, 2024

The topic that I will investigate for this project will be exploring the relationships between health characteristics and behaviors to determine if these factors can predict insurance medical costs.

Business Problem:

Overall annual private medical costs have soared by 55% from 2009 to 2019 with total costs increasing from 709.5B USD to 1064.1B USD. Conversely, medical insurance companies such as United Health have seen tremendous success with a 88.1% increase of gross profits over this same 10 years period (www.macrotrends.net). Since about 66% of Americans are covered by private health insurance (Keisler-Starkey et. al., 2022), it is critical to understand the economic impact of medical costs on the American people especially with the overall health outcomes in the United States declining (www.commonwealthfund.org). Can our choices to have a healthier lifestyle lead to a decrease in medical costs? The purpose of this project will be to investigate if different physical characteristics as well as healthy behavior yield higher medical costs.

Dataset:

The dataset for this project has 1339 rows of private insurance costs along with 6 other demographic and health characteristic information including Age, Sex, BMI, Number of children covered on plan, if the participant is a smoker or not, and region of the United States (Lantz, B., 2019). For this analysis, the insurance costs will be the predictor variable with the remaining variables as independent variables.

Method:

Descriptive statistics and frequencies will be run on all data to ensure there are no outliers and that the distributions are as expected. Exploratory data analysis (EDA) will include different visualizations of these data, line charts of Age x BMI, Age x Costs will be created, and different groupings of the data explored to see if any relationships can be shown visually. Likewise, correlation matrix will be run to determine if what relationships exist and if any variable is not contributing to the analysis. Point-biserial correlations will be explored for both Sex and Smoker variables. For modeling, a linear regression will be conducted using Costs as the dependent variable and all other variables as the independent variables (unless they are removed from the analysis). Other modeling techniques may be explored such as random forest regression.

Ethical considerations:

Ethical considerations for this data include that I was not able to verify the source of this data and thus, it is not clear if this is actual insurance costs or if this data is created. Assuming that the data is real, it is important to interpret any results with caution as there are 1336 rows of data. Other ethical considerations are the number of limited demographics and health characteristics on this dataset. Model results may not be extrapolated to others who share similar characteristics due to the limited number of demographics and health indicators in model development.

Challenges/Issues:

Issues that could arise are that the regression modeling doesn't go as planned and I will have to determine if the data source is viable. Likewise, I could face challenges with interpretability of the data – meaning the sample is not very large. This type of modeling work to

answer what I am interested in exploring would be a much larger project and would take months and possibly years to fully develop a working, robust model.

References:

To validate the results with outside resources, sites like The Commonwealth Fund could be reviewed as well as the US Census, NIH, and CDC sites to see what other research has been done on predicting medical costs using health outcomes. Likewise, I would investigate what publications exist about the costs of private insurance with respect to certain factors like BMI or being a smoker and compare these results to the different results I find with my analyses.

References:

Gross domestic product, national health expenditures, per capita amounts, percent distribution, and average annual percent change: United States, selected yearers 1960–2019 (2021). Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group, National Health Expenditure Accounts, National health expenditures. https://www.cdc.gov/nchs/hus/topics/health-care-expenditures.htm#references

Keisler-Starkey, K., Bunch, L., and Lindstrom, R. A. (2023, September). Health Insurance Coverage in the United States: 2022. Current Population Reports. https://www.census.gov/content/dam/Census/library/publications/2023/demo/p60-281.pdf

Lantz, B. (2019). Machine Learning with R: Expert techniques for predictive modeling. Packt Publishing Ltd.

UnitedHealth Group Gross Profit 2010-2023 | UNH (accessed 2024, April 13). Gross Profits. https://www.macrotrends.net/stocks/charts/UNH/unitedhealth-group/gross-profit

U.S. Health Care from a Global Perspective, 2022: Accelerating Spending, Worsening Outcomes (accessed 2024, April 13). Area of focus: Improving Health Care Quality. The Commonwealth Fund. https://www.commonwealthfund.org/publications/issue-briefs/2023/jan/us-health-care-global-perspective-2022#:~:text=The%20U.S.%20has%20the%20lowest,nearly%20twice%20the%20OECD%20average.