Project 9

Predicting Medical Costs

White Paper

Kristie Kooken

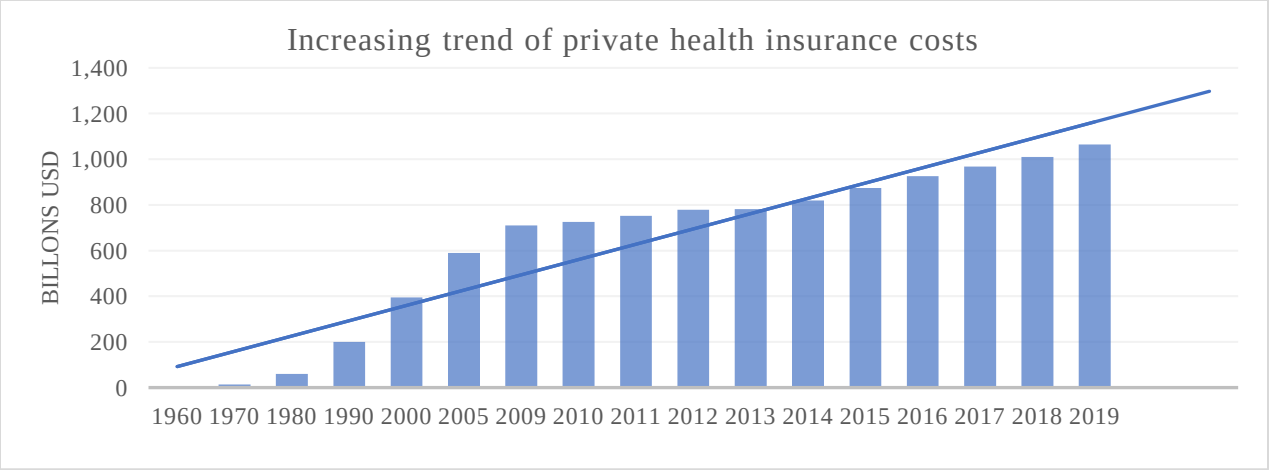
Business Problem

Medical costs create financial and psychological stress in the United States with about three out of four adults stating that they are “very worried” or “somewhat worried” about being able to afford unexpected medical bills (74%) or the cost of health care services (73%) for themselves and their families (www.kff.org). Likewise, about half of adults express that they would not be able to pay an unexpected medical cost of $500 without going into debt (www.kff.org). Can Americans avoid these financial concerns by leading more healthy lifestyles? How do health characteristics influence and predict health care costs? 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.

Background/History

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 (Figure 1). Conversely, medical insurance companies such as United Health have seen tremendous success with an 88.1% increase in gross profits over this same 10-year 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). By investigating what health characteristics can predict medical costs, it would then be possible to develop interventions related to these health characteristics or behaviors.

Figure 1.



Note: www.macrotrends.net

Data Explanation (Data Prep/Data Dictionary)

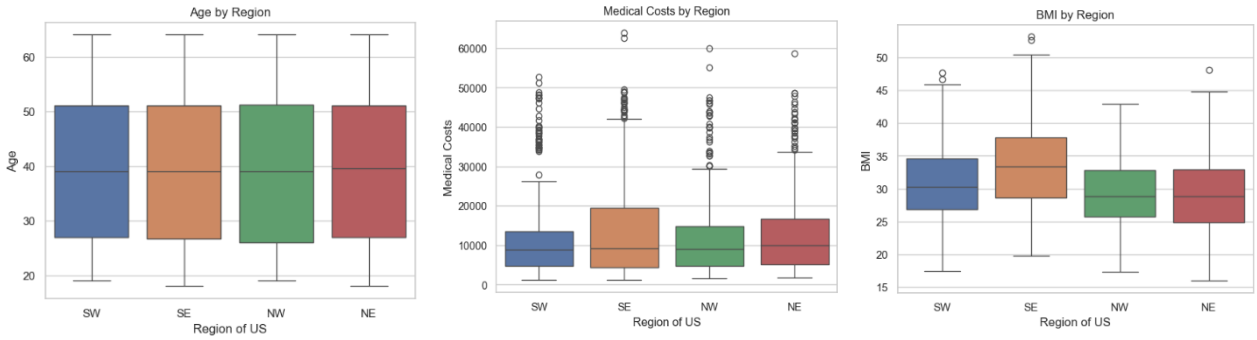
A dataset from kaggle.com was used for analysis that has 1339 rows of private insurance costs along with 6 other demographic and health characteristic information including Age, Sex, BMI, Number of children covered on plan, if the participant is a smoker or not, and region of the United States (Lantz, B., 2019). All data cleaning and analyses were conducted in python using Jupyter Notebook. This dataset consists of 7 columns and the data dictionary for these columns is in Appendix 1.

Methods

This data was scrubbed by ensuring there were no missing values and removing any duplicates. Once this was completed, exploratory data analysis (EDA) was conducted to check the distribution of each variable as well as determine if there were any outliers. Variable distributions were as expected, and no outliers were determined to need further adjustment or exclusion. The resulting dataset had 1337 rows for analysis.

Box plots (Figure 2) were run for each region by Age, Medical Costs and BMI.

Figure 2.

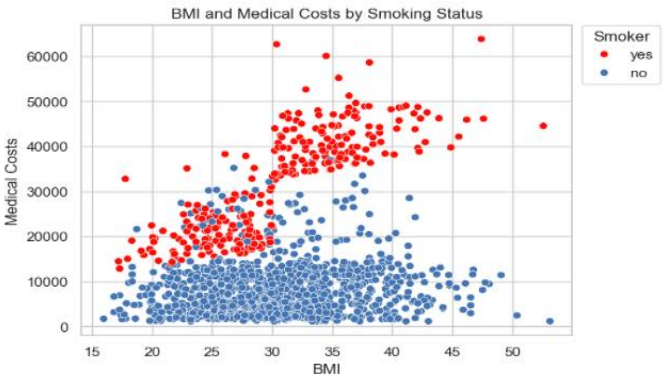


For Medical Costs and BMI, no action was taken for what appear to be potential outliers as medical costs can be extremely high for a particular individual for a variety of reasons and it was assumed that these high costs were accurate. Likewise, with BMI, there were a few cases where BMI values were very high. These were also assumed to be accurate and thus no action was taken on these high values.

Analysis

Correlation analyses were conducted on these data. Results showed that not being a smoker had a significant correlation to lower medical costs (r=-0.7872). Likewise, being a smoker had a significant correlation to an increase in BMI (r=0.8911). All other correlations between Gender, Age, BMI, Number of children covered in plan, region, or Medical Costs were small and insignificant. Interestingly, when looking at BMI by Medical Costs grouped by Smoker status, these data showed that Smoker status appeared to have a strong relationship with these two variables (Figure 3).

Figure 3.



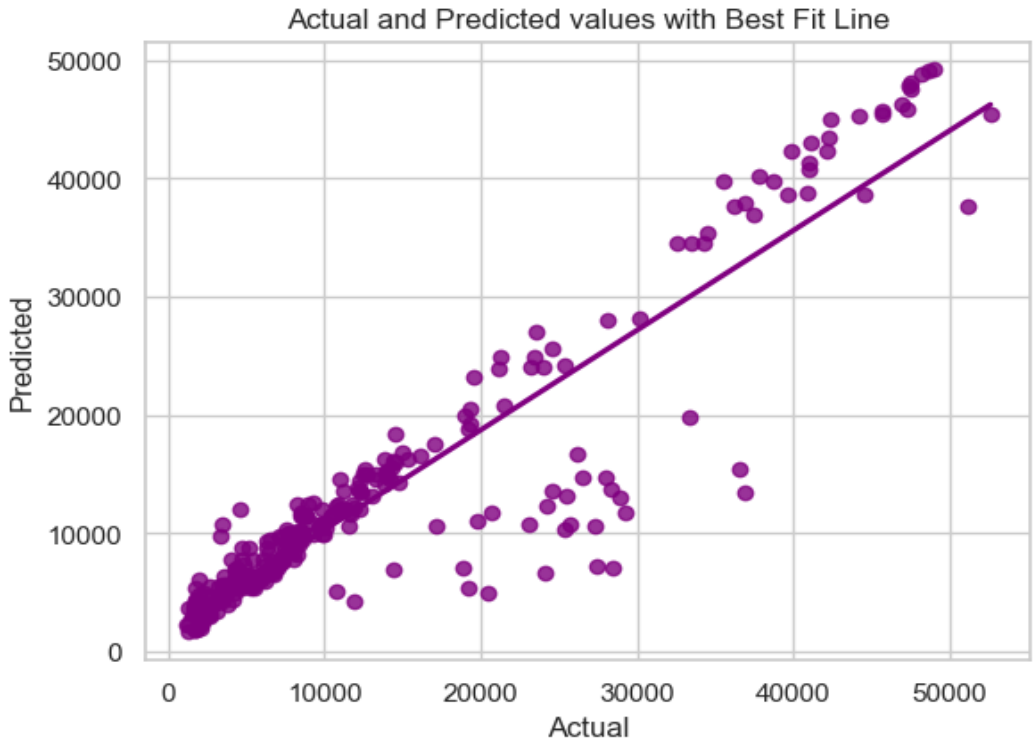
To prepare the data for modeling, those categorical variables with more than 2 levels were dummied to convert each category into a binary numerical variable. Likewise, Gender and Smoking Status were converted to numeric variables. To further investigate the relationships with these data, a linear regression model was conducted to predict medical costs with all other variables as independent variables. This yielded an acceptable model. To explore if another modeling technique could optimize these results, a principal component analysis (PCA) was also conducted to see if reducing the dimensionality of the dataset would yield a stronger model. Model results from this method were also acceptable though the R-squared value was slightly lower. Finally, random forest regression was conducted on these data. This method yielded the best model for this data. Because random forest regression yielded the best method, hyper-tuning was done to optimize model performance. R-squared values and root mean squared errors (RMSE) were generated for each model. The results from these models are presented in Table 1.

Table 1.

Model	R-squared	RMSE
Linear regression	0.7530	6445.635
PCA	0.7232	6824.066
Random forest regression	0.8058	5212.002
Hyper-tuned random forest regression	0.8194	5048.156

Inspection of the independent variables relative importance to predicting medical costs showed that being a smoker had the largest effect (0.679680), followed by BMI (0.178373) and Age (0.119367). Other factors like Number of children and Region had smaller contributions to this model. Figure 4 shows the actual versus predicted values for the final model.

Figure 4.



Conclusion

This project investigated if medical costs can be predicted by an individual's health characteristics and behaviors. Correlation, exploratory analysis, and predictive modeling all demonstrated that smoking is a strong predictor of increased medical costs. Likewise, to a lesser extent, BMI is also a predictor of increased medical costs as well as the age of the patient. As expected with a strong result like smoking, this relationship was seen through the different stages of analyses. The United States has seen a large decrease in cigarette smoking in the past 20 years, with 20.9% of adults smoking in 2005 to 11.5% in 2021 (www.cdc.com). Despite this, smoking has a tremendous economic impact, with smoking-related illnesses costing over 300B USD annually (www.cdc.com). Results from this modeling effort echo this finding, smoking has a tremendous impact on medical costs and an individual can alter potential health costs by not smoking.

Assumptions

A key assumption for this project is that the data presented here is accurate and real medical costs data. Because the data does not have a verifiable source, it will be necessary to recreate any result presented here with additional data either from medical claims or potentially some other government source. Likewise, another assumption is that the large RMSE value can also bring in a potential that these data or this model could make mistakes in prediction and less accurate forecasts. Again, further work in this area is needed in order to determine if it is possible to reduce the RMSE either with more robust data or more refined modeling techniques.

Limitations

Limitations of this work include a smaller sample size of data for modeling. Medical claims data is very large, and it is possible to obtain many gigabytes, if not terabytes, of claims data. This would allow for a much more robust analysis and modeling effort. Other limitations include the fewer number of independent variables in this dataset. It would be important to expand this to as many patient characteristics as well as health behaviors as possible to determine what a final model could include for covariates. Another limitation, which is more of an ethical concern, is that this model identified a relationship between smoking and increased medical costs. However, interpretation must remain very limited unless the underlying data can be verified.

Challenges

Challenges that were faced with this analysis include limited amounts of data, not knowing the source of the data, as well as being unsure if having a large RMSE means that this model will not predict well or if the RMSE for this type of data would be larger.

Future Uses/Additional Applications

Further work is needed to better identify a full set of covariates to predict medical costs. Likewise, it would be ideal for this work to be conducted by the CDC or another government agency so that interventions could be developed to help people who have many health or behavior characteristics that could lead to higher medical costs (instead of an insurance company that would potentially use the findings to increase their revenue streams).

Recommendations

Recommendations include getting more claims data over multiple years to build a more robust model and to have a deeper investigation of the performance of different independent variables included in this model.

Implementation Plan

Currently, there is no implementation plan as this research effort is preliminary and additional data and modeling is needed to develop a robust intervention for health characteristics or behaviors that can lead to higher medical costs.

Ethical Assessment

A huge ethical concern with these types of modeling efforts is likely to be who uses the model. If an insurance company created such a model, then it could be construed as very controversial if they dropped insured individuals based on their health indicators for high medical costs. Other ethical considerations would include ensuring that there was sufficient determination of covariates in order to allow for generalization of results across different communities in the United States.

References

Current Cigarette Smoking Among Adults in the United States (accessed 2024, April 28). Smoking and Tobacco Use. Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm.

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

Health Topics – Tobacco (accessed 2024, April 28). Office of Policy, Performance, and Evaluation, Centers for Disease Control and Prevention. https://www.cdc.gov/policy/polaris/healthtopics/tobacco/index.html#:~:text=Smoking%2Drelated%20illness%20in%20the,due%20to%20secondhand%20smoke%20exposure.

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.

Lopes, L., Montero, A., Presiado, M., Hamel, L. (2024, March 1). Americans' Challenges with Health Care Costs. KFF, The independent source for health policy research, polling, and news. https://www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/#:~:text=About%20three%20in%20four%20adults,for%20themselves%20and%20their%20families.

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.

Appendix 1.

Variable	Decode
Age	Age of primary beneficiary
Sex	Insurance contractor gender, female, male
BMI	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9
Children	Number of children covered by health insurance / Number of dependents
Smoker	Smoking Status, Yes/No
Region	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Charges	Individual medical costs billed by health insurance