Health Insurance Premiums

Lara Mechling, Corrina Hanson, Isaac Liem

Bellevue University

DSC 450 Applied Data Science

Professor Alsaleem

October 2, 2022

Health Insurance Premiums

Table of Contents

Table of Contents 2

Introduction 3

Business Problem 4

Method 4

Results 5

Recommendations and Ethical Considerations 6

Conclusion 7

References 8

Appendix 9

Introduction

The health insurance industry was a thirty-one-billion-dollar industry in 2020 and is a staple amenity in the lives of 240 million Americans (National Association of Insurance Commissioners, 2021). Health insurance provides protection from the costs of different types of medical care. Those who are insured, except in the case of Medicare and Medicaid, typically pay monthly premiums to hold health insurance. These premiums are "the amounts that policyholders pay for health coverage. Policyholders must pay their premiums each month regardless of whether they visit a doctor or use any other healthcare service" (What is Health Insurance Premium?, n.d.).

How does the industry determine what an insured person's premium will be? "By using predictive modelling, the insurers can determine the policy premium for the insured based on their behaviors which are indicated by attributes such as age, BMI (Body Mass Index), smoking habits, number of children etcetera" (Kaur, 2018). Using historical premium data insurance companies can build a predictive model to take in an individual's attributes and estimate a cost for their insurance premiums. Using this model provides the company with a more accurate assessment of the risks involved in providing the insurance and mitigating their costs. "Acquiring a comprehensive understanding of customer behaviors and habits from historical data helps insurers to anticipate future behaviors and provide the right insurance product and policy premium" (Kaur, 2018).

Business Problem

Determining which factors play a key role in determining health insurance premiums allows for a more effective predictive model and will aid the industry in not only providing more accurate insurance costs, but lower premiums. The goal of this project is to determine which attributes play the strongest role in determining health insurance premiums and building a model using these attributes to predict a newly insured individual's premium.

Method

"Machine learning can be defined as the process of teaching a computer system which allows it to make accurate predictions after the data is fed into the model" (Bhardwaj & Anand, 2020). The dataset was used to train and test the model and determine the accuracy of the health insurance predictions.

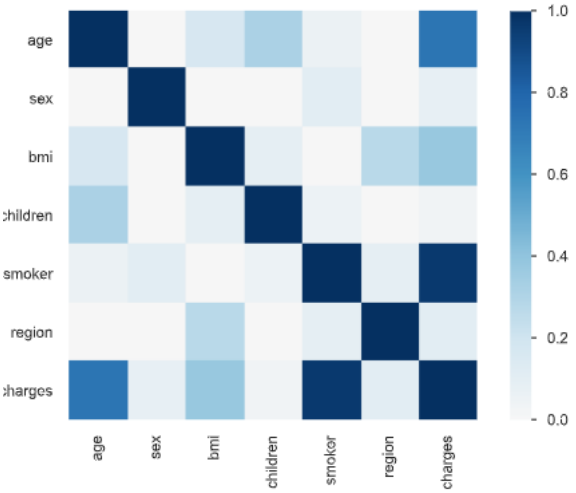
The health insurance premium data was downloaded from Kaggle and imported into Python using Pandas. Two lists were created – a numerical list and a categorical list. Every variable was assigned to one of these lists for ease of use. Exploratory data analysis was performed by visualizing counts and percentages of each variable. A Pandas profile report was also generated for further understanding of the data. One hot encoding was used to create new columns for all categorical variables for feature engineering. All numerical variables were normalized using MinMaxScaler to increase model accuracy.

The target variable for this project is the cost – in the data the column name is 'charges'. In order to create the linear regression model, the x and y variables were assigned. The x variable consists of every data point, excluding the 'charges' column, which was assigned as the y variable. Next the train, test, split function was run with 40% of the data being retained for

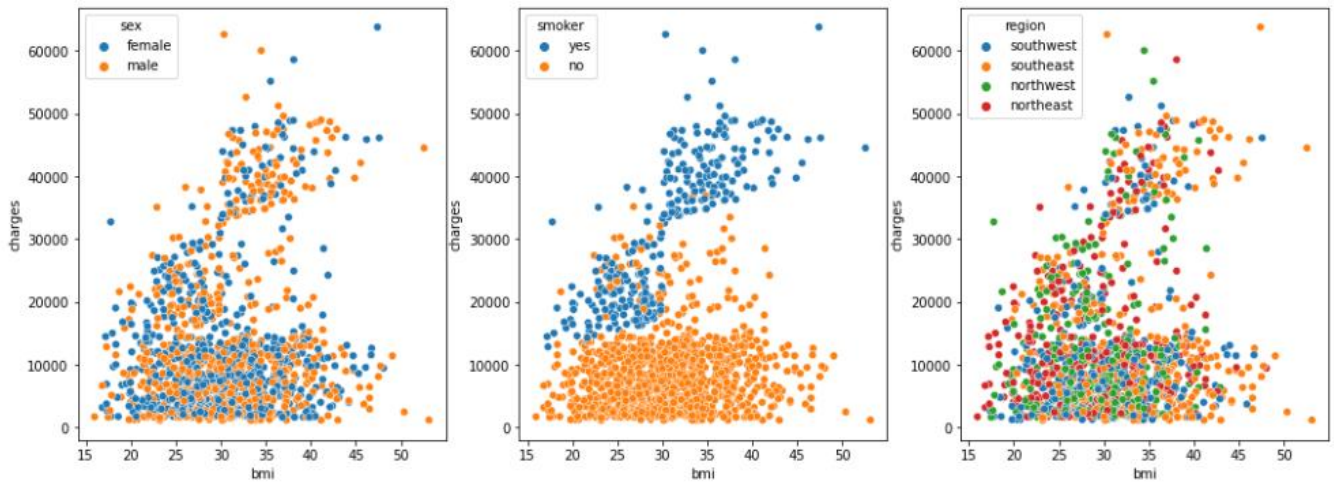
testing, the linear regression model was instantiated, and the model was fit. The coefficients of determination – or R2 scores – were calculated for the train and test data. This was done to measure accuracy. Other figures found for weighing model accuracy were mean square error (MSE) and mean absolute error (MAE).

Results

Correlation plot (or heatmap?) to show relationship/strength of relationship between variables. Age, BMI, and smoker status appear to affect charges the most. The most highly correlated?



Scatter plots of charges by BMI and sex/smoker/region



The coefficient of determination (R2) for the training data was 0.7215 and the coefficient of determination for the testing set was 0.7872. This means that the regression model has an accuracy of 78.72% accuracy.

Recommendations and Ethical Considerations

Using personal health data to predict insurance premiums comes with some ethical considerations. “Because our health insurance landscape currently requires disclosure of a great deal of confidential health information for processing of claims and other administrative purposes, meeting this ethical obligation presents a major challenge, requiring policy solutions that are emerging but not yet fully defined” (English & Lewis, 2016). Health data is protected by governmental HIPPA standards and cannot be disclosed without a patient’s consent and their understand of how the data will be used and shared. With these things in mind, it may be difficult for insurance agencies to gather data on the newly insured clients unless their express written consent is given. Looking at the fact that the predictive model can best align health insurance costs with a person’s medical history, and the ethical standards of how personal health information can be used, the recommendation is for insurance companies to offer voluntary

health questionnaires to prospective clients with the ability to use these questionnaires to reduce their health insurance premiums.

Conclusion

Various factors were used in relation to their effect on health insurance premiums. In alignment with the “Health Insurance Amount Prediction” by Bhardwaj and Anand it was found that smoking status and age had the highest effect on the amount of a premium an insured individual would carry. Using this information and the factors a predictive model was built to predict health insurance premiums for prospective clients. The model was trained and performed with a 78% accuracy on the testing dataset. This “premium amount prediction focuses on a person’s own health rather than … company insurance terms and conditions” (Bhardwaj & Anand, 2020). This information can be used to better align health insurance premiums to the individual creating a more cost-effective model for the insurance company in question. The issue that arises with this is that health data is HIPPA protected and must be disclosed at the behest of those involved. Our recommendation is that health insurance companies use voluntary questionnaires to better evaluate the health of those requesting to be insured in order to come to a middle road in using a model to predict the best premium and still protect client health data.

References

Bhardwaj, N., & Anand, R. (2020). Health Insurance Amount Prediction. *International Journal of Engineering Research and Technology*.

English, A., & Lewis, J. (2016, March). *Privacy Protection in Billing and Insurance Communications*. Retrieved from AMA Journal of Ethics: https://journalofethics.ama-assn.org/article/privacy-protection-billing-and-health-insurance-communications/2016-03

Kaur, T. (2018). Factors Affecting Health Insurance Premiums: Explorative and Predictive Analysis. Retrieved from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dr.lib.iastate.edu/server/api/core/bitstreams/a8729ea4-0ba4-443a-b74d-d5c745470a79/content

National Association of Insurance Commissioners. (2021). *U. S. Health Insurance Industry 2020 Annual Results.* National Association of Insurance Commissioners.

What is Health Insurance Premium? (n.d.). Retrieved from HealthInsurance.org: https://www.healthinsurance.org/glossary/health-insurance-premium/

Appendix

Kaggle Dataset Variables:

Age

Gender

BMI

Number of Children

Smoker

Region

Charges

Dataset Preview:

# age	A sex	# bmi	# children	✓ smoker	A region	# charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47061
32	male	28.88	0	no	northwest	3866.8552
31	female	25.74	0	no	southeast	3756.6216
46	female	33.44	1	no	southeast	8240.5896